The Application of Multiple Modalities to Improve Home Care and Reminder Systems

David Warnock

MSci., University of Glasgow

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Computer Science
College of Science and Engineering
University of Glasgow
June 2013

Abstract

Existing home care technology tends to be pre-programmed systems limited to one or two interaction modalities. This can make them inaccessible to people with sensory impairments and unable to cope with a dynamic and heterogeneous environment such as the home. This thesis presents research that considers how home care technology can be improved through employing multiple visual, aural, tactile and even olfactory interaction methods. A wide range of modalities were tested to gather a better insight into their properties and merits. That information was used to design and construct Dyna-Cue, a prototype multimodal reminder system. Dyna-Cue was designed to use multiple modalities and to switch between them in real time to maintain higher levels of effectiveness and acceptability. The Dyna-Cue prototype was evaluated against other models of reminder delivery and was shown to be an effective and appropriate tool that can help people to manage their time and activities.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Author: David Warnock

Signed: ____________________

Publications

An overview of this thesis work was presented at the CHI Doctoral Consortium in 2012 [210]. Material from Chapter 3 has been published in the proceedings of INTERACT 2011 [212] and the proceedings of Pervasive Health 2011 [209]. Material from Chapter 4 has been published in the proceedings of ICMI 2011 [211], a special edition of the Health Informatics Journal [213], and the proceedings of CHI 2013 [215].[1] Material from Chapter 6 will be presented at British HCI 2013 and published in the conference proceedings [214]. I was the principal author on all listed publications.

[1]This paper won the 'RepliCHI' award for best practice in replicable experimental design.

Acknowledgements

I would like to thank my supervisor Stephen Brewster, without whom this thesis would not have been possible. Stephen afforded me the opportunity to explore things I found interesting while making sure that I would not go too far off course. I would also like to take this opportunity to thank my second supervisor Helen Purchase, who helped to keep my work on track.

I must thank Marilyn McGee-Lennon, who showed infinite patience and was always there to help when I was stuck, confused and stressed. Marilyn is one of nature's 'busy people', who always seems to be working. Despite this, she was always willing to give me her time to proof-read and discuss my work.

There are many people who contributed to my work in small ways. I'd like to thank Maria Wolters, Phil Gray and the other members of the MMH project who gave me feedback on my work. I would like to thank Vicki Hanson and Michael Crabb of the University of Dundee for their assistance in running the older participant group in Chapter 6. I would also like to thank Ken Turner and Claire Maternaghan for giving me feedback, lending me equipment and allowing me to experiment with the HOMER system.

It is important to acknowledge the constant support given to me by my family and friends. I am eternally grateful to my wife Marina, who kept me fed and sane; to my mother Anne, who got me into this mess by bringing home a PC when I was a child; to my father Andy, who was always supportive and tried his best to understand my work; and to all my friends who were understanding, sympathetic and patient as I became increasingly reclusive.

The work presented in this thesis was funded by the EPSRC (Grant `EP/G069387/1`).

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advances in medical science have significantly reduced mortality rates and increased life expectancy; as a result, significantly more people are reaching old age [167, 225]. As people age they are likely to encounter a number of age-related impairments, particularly sight and hearing impairments [201]. There is also a much higher chance of developing cognitive impairments, in particular problems relating to memory [169, 179, 188]. Assisted living technologies (ALTs) can be deployed to help counteract some of these effects, such as by providing reminders or prompts to help a person manage their environment or activities.

A home reminder system is a form of assistive technology designed to help people living in their own homes. In general, the aim of a home reminder system is to help ensure a high level of autonomy and a high quality of living for the care recipient. It attempts to achieve this by delivering reminders (also known as notifications, prompts and alerts) to its users in order to help them manage their personal care, environment and lifestyle.

There are many challenges when developing such technology. Each home presents a unique environment, requiring a customised solution. Users will each have their own preferences, care needs and impairments, requiring further customisation. The home is often a complex social environment, potentially with multiple inhabitants (including pets). Perhaps most importantly however, both the home and the users are likely to change over time. Users might develop new care needs as they age, and existing age-related issues are likely to worsen. If technology for the home is not built to be flexible, then errors could be produced by small environmental changes such as the presence of guests. Larger changes such as moving house or redecorating could seriously impair the functionality of technology that has not been specifically designed with flexibility in mind.

This thesis considers these problems and outlines two potential solutions. The first is *multimodality*, which in this context means providing several ways to deliver reminders to the user. The second approach is the creation of *dynamic technology*, able to control *when* and *how* communication takes place for more appropriate and effective interactions. These methods can be used together to make flexible and robust home care technology with the ability to compensate with changeable environments and users.

1.1 Motivation

This research is important due to the significant increase in the older population [167] that is occurring throughout the world. Due to the natural effects of ageing, an older population will inevitably mean larger percentage of the population requiring care to maintain a reasonable standard of living.

> *"In common with other European countries, the UK has an ageing population. The proportion of people aged 65 and over is projected to increase from 16% in 2008 to 23% by 2033. This is an inevitable consequence of the age structure of the population alive today..."*
>
> — UK 'Age of the Population' Survey [167]

The shifting population distribution creates two problems, the first of which is a logistical issue: there will simply not be enough care workers to adequately care for the larger aged population. The second problem is an economical one, in that the working population will be significantly diminished. The economical problem is exacerbated in countries which provide a state/social pension[1] and in countries that provide universal healthcare (such as the UK). Countries which provide both (which includes most of Europe) will be required to pay a pension and provide potentially expensive medical care for a sizeable portion of the population.

[1] Almost every 'developed country' in the world provides a state pension, although Germany is an unusual exception. Source: http://www.pension-watch.net.

> *"Despite the forthcoming increases in state pension age, the ratio of working age people to each person of state pensionable age is projected to decline from 3.2 in 2008 to 2.8 in 2033."*
>
> — UK 'Age of the Population' Survey FAQ [166]

A number of measures have been implemented to tackle this, such as changing the rules for pensions. In the UK the default retirement age, previously 65, was removed entirely to encourage some portion of the aged population to remain economically active. In 2010, legislation to increase the French retirement age from 60 to 62 resulted in widespread protests, strikes and blockades. Such measures might help to address the economic problem, but do little to help with the logistical issue; there will simply not be enough carers to provide an adequate level of care. A recent UK study discovered that this is already a serious issue.

> *"...there were many instances of home care which caused us real concern, where human rights were breached or put at risk because of the way care was delivered."*
>
> — Close to Home Report [60, p. 96]

The Close to Home Report [60] highlighted that the quality of home care in the UK was so poor that it often violated the human rights of the patients. The report highlighted ageism, lack of investment, poor training and a lack of communication between carers and patients as being behind the issues. One of the most important issues highlighted in the report was that carers are poorly trained and paid, which prevents the industry from developing and retaining skilled care workers.

> *"Low pay and stretched unpaid hours correlate directly to the high rate of turnover in the workforce. The lower the pay, the higher the churn and turnover rate in the workforce. This benefits no one, making well-trained, person-centred care extremely difficult to deliver."*
>
> — Close to Home Report, [60, p. 74]

An alternative to providing care at home is to move care recipients into care homes (sometimes called residential or institutional care). This would allow carers to increase the number of patients in their charge, while also providing a more uniform service. This might seem at first to be the answer to the economic and logistic problems: unfortunately, this is simply not the case. Kane [113] noted that regardless of the quality of care, care homes generally provided a low quality of living for patients. The My Home Life Report [194] even found that the average living space provided to care home inhabitants was smaller than an average prison cell. The economical argument for residential care homes is also invalid; Chappell *et al.* [44] found that providing home care is "significantly less costly" than residential care. Residential care also creates additional problems, such as the stress involved in removing a person from their own home and placing them in an unfamiliar environment. Many older people might be against moving into residential care, but the decision will be made on their behalf by carers or family members with concerns about their ability to live on their own.

> *"I didn't make any decision about coming here. I was forced here. My flat was sold and then it all went haywire. . . "*
>
> — My Home Life Report [194, p. 47]

Some patients require round-the-clock care, and for those patients residential care may be in their best interests. For other patients with less severe care needs, there is a very strong case for keeping them at home and independent for as long as possible; it results in a higher quality of life [60, 113] and reduces costs for patients and governments [44]. The main issue, as highlighted by the Close to Home Report [60] is building a care network with the skills and personnel needed to provide high quality care. Home care technology can alleviate pressure on such networks by enabling home care recipients to manage their own care, environment and lifestyle. The work presented in this thesis aims to assist in addressing this important real-world problem.

The primary issues which face home care technology will be fully explored in Chapter 2. In summary, the main issues are:

- the needs of home care recipients are highly heterogeneous [60];
- the users of home care technology often have sensory, cognitive and physical impairments [201];
- the needs of home care recipients will change over time [138, 207];
- the home is a challenging and dynamic environment for technology [57].

Users of home care technology are very likely to have one or more sensory impairments such as sight or hearing loss [201]. Multimodal interaction uses one or more of the senses in order to convey information; as such, it is ideal for creating systems for people with sensory impairments. The most commonly encountered interaction modes are visual, aural and tactile; a typical mobile phone comes equipped with the hardware to deliver all three. However, existing home care technology both in research and industry (which will be explored in Sections 2.3 and 2.2 respectively) tends to consider only one or two modalities.

Multiple modalities in home care technology could help to overcome the common issue of sensory impairment by allowing for the delivery of information in non-impaired senses. Multiple modalities could also help address another issue: that home care technology must be flexible enough to cope with evolving care needs and the dynamic environment of the home.

1.2 Thesis Questions

The aim of this thesis was to address the following research questions:

Thesis Question 1:
Which forms of interaction are appropriate for use in a home care system?

Thesis Question 2:
How do different forms of notification delivery affect users?

Thesis Question 3:
How can home reminder technology be designed to best utilise multiple types of interaction?

Thesis Question 4:
Can home reminder technology be made more effective and appropriate by providing it with the ability to dynamically select from multiple forms of interaction?

In this thesis these research questions will be broken down into smaller and more direct research questions. To avoid ambiguity, overarching research questions will be called *thesis questions*, while the sub-questions will simply be called *research questions*.

The overall aim of this work is to support the development of technology that will assist people requiring regular care at home. This will be done by examining additional methods of interaction and reminder delivery to achieve more effective technology that produces a higher level of user satisfaction. It is hoped that this work will be able to assist in tackling the economic and logistical problems of providing a good standard of care for an ageing population.

1.3 Thesis Structure

Chapter 2 of this thesis explores the existing literature in the field to provide an overview of existing work into both home care technology and notification technology, providing an overview of existing guidelines and identifying where further work is needed. Chapter 3 presents the first study of this thesis, which examined eight different notification modalities to explore which ones were the most effective and disruptive when delivering information. Chapter 4 continues this work by examining the differences between distractions and useful notifications with both younger and older users, with the younger users both providing a useful baseline for performance and potentially providing an insight into the abilities of younger care recipients. Chapter 5 introduces Dyna-Cue, a prototype multimodal reminder system with logical rules based on the findings of the first two studies. Chapter 6 describes how a novel task with a home-based context was developed and used to evaluate the Dyna-Cue prototype, comparing it to alternative models of reminder delivery. Chapter 7 summarises, discusses and explores the contributions of this work before concluding the thesis.

Chapter 2

Literature Review

This chapter presents a review of literature relevant to this thesis. The thesis questions set out in Section 1.2 consider ways to improve home care technology. To fully understand how this can be done, a good understanding of existing home care technology is needed. At the root of this is an understanding of how home care technology *should* be designed based on the needs of its users; this is examined in Section 2.1. The current state of the home care industry is explored in Section 2.2 to reveal how well industry offerings meet those requirements, which is a crucial step in understanding how the technology can be improved. This is followed by Section 2.3, in which the home care technology projects in research are reviewed. As this thesis considers reminder technology, Section 2.4 reviews literature on notifications, interruptions and interaction methods to clarify the work that needs to be carried out to address the thesis questions presented in Section 1.2. Finally, this chapter concludes in Section 2.5 with a summary of the main points and a clarification of the research gap that will be addressed by this thesis.

2.1 Requirements for Home Care Technology

The aim of this section is to identify the underlying challenges faced by home care technology. Understanding these challenges is an important step in addressing the thesis questions presented in Section 1.2, and allows critical insight into the value and validity of current home care technology in industry and research. Section 2.1.1 identifies the end users and stakeholders for home care technology, along with their requirements and expectations. Section 2.1.2 outlines some of the problems that are likely to arise when trying to identify the requirements of home care technology. Section 2.1.3 explores the design issues that arise when developing technology for a home environment. Finally, Section 2.1.4 summarises and discusses the points made in this section.

2.1.1 The Users of Home Care Technology

The primary users of home care technology are older people; for example, a report on the users of home care in Scotland found that 80% of all home care service users in 2007 were over the age of 65 (56,701 people in total) [109, p. 5]. The remaining 20% comprised people with mental health problems, learning disabilities and people with physical impairments. This report also states that the older the person, the more likely they are to require care.

> *"The need for [home care] services increases significantly with age, long-term conditions and proximity to death."*
>
> — JIT Report on Home Care Statistics [109, p. 9]

This figure outlines an important question: why do older people make up such a large percentage of care recipients? At a high level, the answer is that as we age our bodies start to fail due to a process called senescence: the gradual failure of a biological system over time. The study of this process in humans is called gerontology, but lies outside the scope of this thesis and will not be discussed further here. However, it is important to understand why so many older people require care services.

One of the reasons for the high demand for care services in older personolder people is the prevalence of sensory impairment. The human body includes a number of different sensory systems, some of which are more resilient than others to the effects of ageing. However, eyesight and hearing problems become increasingly common as we age, with a very high volume of older persons having a sight or hearing impairment.

> *"As well as chronic illness, older people are also more likely to have a disability. Nearly half of disabled people are aged 65 or older. The most common problems relate to movement and to vision and hearing. Sensory impairments become increasingly common as people age: around 80% of people over 60 have a visual impairment, 75% of people over 60 have a hearing impairment, and 22% have both a visual and hearing impairment."*

— National Service Framework for Older People Report [201]

Most aural impairment is due to presbycusis, in which hearing loss develops naturally with age due to the degeneration of fine hairs in the ear. Electronic hearing aids are often used to counteract the symptoms of hearing impairment. Visual impairments can be caused by a number of things including age-related macular degeneration, Cataracts, Myopia (short-sightedness) and Glaucoma. Some conditions like Myopia can be tackled with corrective glasses, while other conditions such as Cataracts are treated surgically.

Other sensory systems are generally more resilient. The sensory receptors in our skin are distributed and heterogeneous; their signals are automatically combined to produce the senses of pain, pressure and temperature. This is called the somatosensory system, and it is quite hardy due to the massive redundancy of the receptors. Somatosensory impairment is generally quite rare and is usually caused by neurological disorders; it is not expected to naturally decline with age.

The senses of taste and smell (gustatory perception and olfactory perception, respectively) have been shown to decline with age, but the mechanism for this is not well understood [28, 114]. It has been suggested olfactory impairment is rarer than gustatory impairment [28], but also that the two are linked and will decline together. Complete impairment of the olfactory or gustatory senses is quite rare but can have a significant impact on the mental well-being of older people [28].

There is a lot of technology that can help with age-related impairments and mobility problems. However, it is much harder to develop technology to tackle mental health issues relating to old age. Along with a much higher chance of dementia [33], most (but not all) types of memory decline with age [169]. One of the most common types of memory failure with age is prospective memory: the part of memory used to remember to do things (sometimes called ‘remembering to remember’) [41, 188]. Working memory is also known to decline with age [179].

These issues explain why such a high volume of home care users are older people. They also highlight some important design issues for home care technology:

- Home care users are likely to have a range of sensory impairments and/or memory problems of various levels of severity.

- Home care users are likely to find that their conditions will worsen as they age and that new conditions will develop.

These issues make it challenging to create a single home care solution that will help a wide range of people. Unfortunately, along with the physical and practical issues involved when designing technology for older users, the design of home care technology is further complicated by the fact that the technology is intended to provide care and that it is intended for the home.

2.1.2 Designing Care Technology

Care technology is a difficult design scenario for a number of reasons. As shown in Section 2.1.1, the needs of those who require care are extremely varied. In addition, there are several groups of stakeholders involved in providing care, including patients, doctors, family, nurses and social workers. This makes it extremely difficult to accurately identify the requirements for home care technology. When considering care technology for the home, it is important to consider the home itself, including how the design will affect other residents. For these reasons, McGee-Lennon argued that traditional methods of requirements capture are not suitable for use in a home care scenario [142].

Various ways have been suggested to tackle these issues. Morgan *et al.* [156] described how to use a theatre piece to reconcile the views of different user groups and stakeholders. Blythe & Dearden [24] suggested *personas*, fictional representations of older persons in various scenarios, to help guide the design progress and provide an accessible insight into the needs of the user for stakeholders who might otherwise not have one. Other researchers such as Demirbilek & Demirkan [53] advocate participatory design; involving the users in the design of care technology. Intille *et al.* [101] suggested storyboards could be used to gather information that would normally be missed, in particular user preferences. Resolving stakeholder conflict lies outside the scope of this thesis; this thesis focusses on ways that the problems outlined by McGee-Lennon [142], including the shortcomings of inadequate requirements capture, can be addressed at the design stage by creating more adaptable and flexible technology.

Fickas [62] may have a solution in ‘clinical requirements engineering’: the idea that you can reconcile clinical trial methods with requirements engineering through iterative development. In Fickas’ work, participants with cognitive impairments were asked to use an e-mail application. All participants started with the same application and their own set of goals. Each participant was observed to identify unique problems they were having with the base application. The base application was then reprogrammed for each participant individually, addressing their specific problems. This had two outcomes; firstly, multiple different final versions of the e-mail application were developed; and secondly, many of the participants reached their self-set goals. Unfortunately, this method is very labour-intensive, making it an impractical way to develop home care technologies for general use.

The clinical requirements engineering approach could implemented if technology was designed to be more adaptable, especially if this can be done by the users themselves. Wang *et al.* [208] outlined how this could be done with a policy system; software that uses a set of rules to define its behaviour. McBryan *et al.* [137] showed how such technology could be used to allow evolution in a home care system, and Wang & Turner [206] showed how such a system could be used to resolve stakeholder conflicts. The existing work here suggests that, due to the difficulties in performing a requirements capture for home care technology, the technology should not be generalised but instead made to be adaptable so that it can evolve iteratively.

Adaptability is therefore an important requirement for home care technology. Wang & Turner [207] identified adaptability as one of four key concepts that home care systems should always consider along with personalisation, customisation and dependability. Personalisation refers to creating a system that suits the user's needs and customisation means the system can be tailored to suit the user without the aid of an engineer.

Surprisingly, Wang & Turner's [207] work does not also consider how to make home care technology acceptable for the users or their home. If users reject care technology, then they might ignore or disable it as a result. For example, McBryan *et al.* [137] described a scenario where a user disabled notifications by turning off his phone because they were annoying. Such an outcome would be a critical failure in any piece of technology, but the result of this could be disastrous for technology designed to provide assistance or care. Therefore, it is vital that such technology is accepted by the users; therefore, it is necessary to understand what makes technology acceptable in the home.

2.1.3 Designing for the Home

Mark Weiser's seminal article "The Computer for the 21st Century" [217] was published in 1991. The article described his vision of Ubiquitous Computing (UbiComp): technology that would be so "so ubiquitous, that no-one will notice [its] presence" [217, p. 19]. Weiser gave an example of a UbiComp world where the home, car and workplace are all interconnected with small inexpensive computers. The interactions he described taking place in the home are interesting; the alarm clock offered coffee, but only understood 'yes' and 'no' as answers. Despite the ubiquity of cheap display screens, Weiser envisioned that people would still read newspapers printed on paper; yet the protagonist has a pen able to digitise an article and send it to her office. The final example given is a piece of technology the protagonist does not know how to use: a device to track down a lost instruction book. Even in Weiser's future where technology is inexpensive and ubiquitous, its role in home life is unclear.

Weiser's vision for UbiComp has not (yet) come to pass. However, Bell & Dourish [19] argue we are now living in a UbiComp future, pointing to the ubiquity of internet-connected devices in Korea and Singapore as an example. Bell & Dourish criticise Weiser's vision for being an 'American' view point and, in their study of Korea, note that Koreans with home computers will often still pay to use public Internet cafés due to the culture surrounding the home. If Bell & Dourish are correct then we are currently living in a UbiComp world where, primarily for cultural reasons, technology in the home is not being used by the residents.

> *"...Korean homes are considered to be extremely private domains, closed often even to one's closest friends, and that socializing, especially when it comes to gaming, has nearly always had a space in the public domain, and is in fact actively sought out that way. While 90% of Korea's online population goes on line at home, 25% also report regularly going online in cyber cafés."*
>
> — Bell *et al.* [19, p. 138]

Developing technology for the home is complicated by a number of factors. Different homes will have different layouts, furniture, decorations and room sizes. There may be multiple residents and different species of pet. Domestic routines will vary enormously, and can be expected to change in an unpredictable way. The environment itself is also likely to change with time, from large things such as redecoration to small things like moving a chair for guests. Technology for the home is also at the mercy of a number of subjective factors such as aesthetics, perceptions of value or worth (*e.g.* the 'cool factor' of a gadget) and personal preference. With so many factors at play, how can technology be designed to be acceptable in the home?

Edwards & Grinter [57] argue that UbiComp technology aimed at the home has seven challenges to overcome. First, that existing homes have not been designed with smart technology in mind. They theorise that most users will upgrade their home piece by piece, as opposed to completely outfitting a home with the required technology.

The second challenge is closely related to the first; that home care technology is likely to be a heterogeneous evolution. Newer devices should interact with older ones without problems. The authors note that an old rotary phone will still work with a modern phone connection, and argue that ubiquitous computing devices for the home should show a similar level of robustness.

Their third challenge is that UbiComp technology for the home must be expected to work without the need for technical expertise. Most home technology, such as televisions and washing machines, are designed to require minimal technical expertise to use. This is vitally important when potential users include elderly users who may have no technical skills.

The fourth challenge is that such technology should be aware of its potential impact on domestic routines. For example, if person A watches the daily news at 6pm, then addition of 'Sky+' or 'TiVo'[1] might change this routine by allowing the person to record the news to watch at 7pm instead. This is a particularly complex issue and difficult to predict, as domestic routines vary from household to household.

The fifth challenge is the social impact of such technology. Unfortunately, this is usually an emergent property that is difficult to establish beforehand. An obvious example is the Internet, which has had a deep reaching social impact that was impossible to predict at its creation.

The sixth challenge, which is of utmost importance to any home care technology, is reliability. Components of the system must be significantly more reliable than current computer systems; Edward & Grinter argue that it should be as reliable as the telephone. It should also be able to degrade gracefully, compensating for failure wherever possible.

The seventh and final challenge regards how much a UbiComp system can reliably infer from limited data: *i.e.* how context-aware the technology is. This is one of the hardest things for home technology to achieve due to the complexity and variability found between homes and residents.

[1] Sky+ and TiVo are Digital Video Recording (DVR) devices, capable of automatically recording a live broadcast for replay at a later time.

Crabtree *et al.* [50] suggested that the best way for UbiComp technology to move from the workplace into the home would be to try and support everyday activities, in particular communication between residents in the home [49]. Crabtree & Rodden [48] suggested that ethnographic studies can be used to identify domestic routines and the areas of the home which are used to facilitate communication (*e.g.* tables, fridge doors) which would reveal the best places to target home technology.

Taylor *et al.* [193] embodied this principle with their augmented fridge magnets, which aimed to support an existing domestic routine: the use of the fridge as an inter-resident communication platform. Taylor *et al.* concluded that their lightweight solution was effective because it was lightweight, but also pointed out that several lightweight systems could eventually work together to support domestic routines. Another example of this is the InPhase system by Tsujita *et al.* [199], which uses domestic routines to facilitate communication between the residents of separate households.

O'Brien *et al.* [165] carried out an ethnographic study and concluded that technology for the home would not only need to embrace domestic routines, but it would also need to be highly flexible to cope with the differences between household and the way homes change over time. Adaptability is a recurring theme for home care technology: Wang & Turner [207] also argued that home care technology would need to be adaptable, but primarily to deal with the dynamic requirements encountered when making home care technology. Edwards & Grinter [57] argued that the 'smart home' would evolve over time, and would need to be able to adapt to suit.

2.1.4 Observations on Home Care Requirements

In Section 1.2, Thesis Question 1 considered which forms of interaction are suitable for use in the home, while Thesis Question 4 considered how home care technology could be made more effective and appropriate. Creating an effective home care system requires the correctly identification of the requirements for that technology. The literature discussed in this section reveals a number of issues that complicate the process of capturing those requirements, such as the high likelihood of sensory impairment and situational changes that would necessitate changes to the system.

Most of these problems cannot accurately be predicted at the requirements capture stage, and as such a common theme in the literature is that home care technology must be adaptable enough to compensate when the requirements change [57, 62, 137, 207]. Ideally, adaptation should not be a process that requires an engineer or system upgrades;

the home care solution itself should be programmable and able to interact with any new technology that becomes available.

There are also other requirements for such technology. In order to remain acceptable in the home, it must respect the home environment. One way to achieve this is to ensure it does not disrupt existing domestic routines; in the example given by McBryan *et al.* [137], the technology disrupts a domestic routine (watching TV) and is disabled as a result. Home care technology must be personalised to suit the unique requirements of the user, in particular considering things such as sensory deprivation and memory issues. The technology must also be reliable, especially if will be charged with care.

The existing work suggests that flexibility is a key component of making effective home care technology, providing an initial insight into how thesis questions 3 and 4 can be answered. This section has also highlighted the need for alternative methods of interaction to overcome sensory impairments; with respect to Thesis Question 1, it's clear that limiting home care technology to visual and audio interactions would limit the effectiveness of the technology. The following section examines the products currently offered by the home care industry to evaluate how they have addressed these challenges.

2.2 The Home Care Industry

Thesis question 4 considers how home care technology could be made more appropriate and effective. There are many types of assistive technology available and a range of companies that research and develop their own products. An understanding of those products must be formed to fully answer Thesis Question 4, allowing areas for improvement to be outlined and to identify existing products which work well.

Although there is a great deal of research on different types of home care technology, there is no *de facto* taxonomy, and sometimes literature will use terms such as 'telehealth' to refer to very different things. The Disabled Living Foundation is a British charity that offers advice and assistance for people involved in home care, and they provide certification services for health and social care workers. Their advice website, Living Made Easy [131], lists five categories for assistive technology: telecare, telehealth, 'simple solutions', activity monitoring and environment management. This section explores each of these terms in Sections 2.2.1 to 2.2.5, each of which provides a definition and overview of that type of home care technology. Section 2.2.6 summarises the main points made in this section.

Note that the terms *telecare, telehealth* and *telehealthcare* are often used interchangeably by both research and industry. At present there is no standard taxonomy or terminology for this type of assistive technology, and that some types of technology do not fit neatly within the categories identified by the Disabled Living Foundation [131].

2.2.1 Telecare

> *"Telecare consists of equipment and services that support your safety and independence in your own home. The equipment can sense risks such as smoke, floods and gas, can remind you to take pills and even call for help if you fall."*
>
> — Living Made Easy [131]

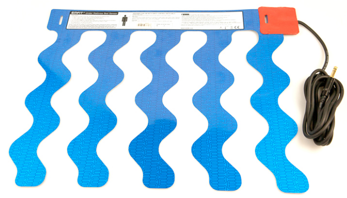
Tunstall[2] is one of the most prolific manufacturers of this type of technology. They produce a wide range of devices including fall detectors, flood detectors, gas detectors, panic buttons, carbon monoxide detectors and seizure detectors, some of which are shown in Figure 2.1. For interacting with the user, the Tunstall catalogue [200] includes pagers, visual alarms, audio alarms and X10 controllers. However, it is heavily implied that these devices are for communication with a carer or family member and not the telecare user. Other companies including Chubb Community Care,[3] Tynetec Ltd.[4] and CareTech UK[5] also manufacture these types of sensors, with little to differentiate between each of the company's products.

Due to the variety of sensors available and the modular model used by manufacturers, telecare technology solutions can be constructed to address specific problems, *e.g.* a fall detection system can be built for someone at a higher risk of falling. These systems can help to ensure the safety of a care subject without constant supervision, which in turn reduces the pressure on families and carers. The modular nature of these systems also means that they can adjust with changing care needs and environmental changes, although such changes are likely to require an engineer visiting the home.

[2] Tunstall: `http://www.tunstall.com`.
[3] Chubb Community Care: `http://www.chubbcommunitycare.co.uk/`
[4] Tynetec Ltd: `http://www.tynetec.co.uk/`
[5] CareTech UK Ltd: `http://www.caretech.co.uk/`

(a) Flood Detector

(b) CO Detector



(c) Seizure Sensor

(d) Fall Detector

(e) Natural Gas Detector

(f) Bed Use Sensor

Figure 2.1: A selection of sensors manufactured by Tunstall. Images ©Tunstall.

While current-generation telecare systems might be very effective at reassuring carer and family, the opinions of the care recipient do not appear to be a priority for developers. The systems are designed to detect emergency situations and alert carer or family, so although the technology is placed in the resident's home they have little or no control over it. As can be seen in Figure 2.1, the system hardware is generally not designed to blend with the aesthetics of the home. Although it would be difficult to design technology that would please everyone, little to no attempt is made to make the hardware aesthetically pleasing and no customisation options are offered. This could leave residents feeling as though they have little control over their own care, which was a common theme in the Close to Home Report: "*Many felt that home care was something that was being 'done to them', rather than a service that they could engage with and arrange to suit their lives and needs.*" [60, p. 57].

It is likely that a higher standard of living for the user could be reached by making the user the focus of this technology, instead of their family or carers. For example, simple reminder messages could be used to interact directly with the user to promote healthy daily routines, ensuring a greater degree of independence. Although the Living Made Easy [131] definition specifically refers to reminders, technology that actually provides this functionality is rare; the best example is the Tunstall LifeLine Vi Plus, a personal alarm with programmable speech reminders.

The lack of aesthetic flexibility highlights that user preferences[6] are at best a secondary concern, yet Section 2.1 shows that users expect technology to respect their home and daily routines. While this technology is flexible, it does not allow for user preferences to be taken into account. This is likely to have a negative effect on how appropriate the technology is for the home, and has clear implications for Thesis Question 4 which considers how home care technology could be improved.

2.2.2 Telehealth

> *"Telehealth covers the electronic exchange of personal health data from a patient at home to medical staff at hospital or similar site to assist in diagnosis and monitoring."*
>
> — Living Made Easy [131]

Telehealth products generally provide the following functions: *(1)* to assist a person in self-management or monitoring of medical conditions and *(2)* to provide richer communication between the user and medical professionals or carers. This sort of technology is generally deployed in the homes of people with heart problems, diabetes or similar medical conditions that require a higher level of observation, providing a greater degree of independence without lowering the quality of their medical treatment. The term telehealth has also been used to describe technology that allows medical services to be provided to rural and isolated areas, *e.g.* via secure video conferencing. As this thesis focusses on home care technology only telehealth monitoring devices will be discussed.

Along with telecare technology, Tunstall also manufacture telehealth monitoring devices. Tunstall products can be combined to create a hybrid telecare/telehealth system. The Tunstall 'MyMedic' product includes a colour display, audio announcements and large buttons for ease of use. As shown in Figure 2.2a, the device looks somewhat like an

[6]Throughout this thesis the 'user' of home care technology is given to be the resident care recipient, although there are often many people involved in the operation of home care technology. This is discussed further in Section 2.1.1.

oversized pager. Along with information provided by the user, this device is able to monitor sensor data for a better overview of the user's health.

As with telecare products there is a range of products on offer; however, there is more variety between telehealth products. The Decobo HealthHUB^{TM},[7] shown in Figure 2.2b, is slightly more complex than the Tunstall MyMedic and is equiped with a larger monochrome screen. Other products such as the 'Home Pod' by Telehealth Solutions[8] and the Honeywell HomMed[9] system are actually software products which can be run on a PC or Tablet Computer.

Existing telehealth technology has generally not been designed for people with cognitive, sensory or physical impairments. Yet when it comes to home care technology users, Section 2.1 has made it clear that users are likely to have one or more sensory or cognitive impairments. Of the 4 solutions discussed, only the Tunstall MyMedic appears built with this in mind; the brochure [200] draws attention to the multiple interaction modes on offer, the size of the buttons and the simplicity of the device's interface. Multimodal interaction can be used to overcome sensory impairment, giving the MyMedic a distinct advantage over its competitors. More technology should embrace multimodal interaction to promote higher levels of accessibility.

These products are important as they involve the end user in their own care. While many of them could be criticised for neglecting users with special needs, the variety of devices on the market suggest that many telecare users would be able to find a device which meets their unique requirements.

2.2.3 Simple Solutions

'Simple' care solutions are dedicated standalone products that address a specific need. Examples include medicine dispensers, automatic lights, big-button phones and simplified remote controllers. The sensor devices used in telecare, such as the ones shown in Figure 2.1, often have a 'simple' counterpart that activates a built-in alarm when the sensor is triggered, *e.g.* a fire alarm. Such devices can be more affordable than telecare technology, but as they are standalone they are unable to communicate emergency situations to other parties. While such devices might form a useful part of care at home, the fact that they are standalone makes them of little interest when performing

[7] HealthHUB^{TM}by Doboco Ltd, `http://www.docobo.co.uk`.
[8] Health Check Pod by Telehealth Solutions, `http://www.telehealthsolutions.co.uk/`.
[9] Honeywell HomMed, `http://www.hommed.com/`.

(a) Tunstall MyMedic Telehealth Device. Image ©Tunstall.

(b) Doboco HealthHUB^{TM}. Image ©Doboco Ltd.

Figure 2.2: A comparison of dedicatedtelehealth monitoring devices from two manufacturers.

research into comprehensive home care technologies. For this reason such devices will not be discussed further in this thesis.

2.2.4 Activity Monitoring

> *"Activity monitoring systems monitor your daily activities in your home and provide a chart of your activity via the internet. Small, wireless sensors are triggered as you move around your home. A carer or relative can therefore see if you have not gone into your kitchen all day, or have gone out and not returned."*
>
> — Living Made Easy [131]

Activity monitoring technology overlaps somewhat with both telecare and telehealth. Such systems will use an array of sensors to monitor a care recipient not just to detect emergency situations, but also to collect general lifestyle measures such as the time spent sleeping or in the bathroom. While the care recipient might be able to interact with the system in some ways, existing technology is almost always based on a sensor network that provides information only for carers. Examples of such technology include wandering detectors (to detect if a user has left home at an unusual time or has not returned), bed-use sensors (such as the one shown in Figure 2.1f), room occupancy sensors and similar devices [200].

One example of such a device is the ‘Just Checking Carer’ by JustChecking Ltd.[10] This is a wireless system that uses sensors to record the movement of a person around their own home. Movements are plotted on an online activity chart which other stakeholders can use to check the daily activity of the care recipient. Like the telecare technology discussed before, the product has a base unit and several wireless sensors, all in white plastic, with no communication to the resident. Therefore, care recipients might not see any direct benefit of this technology and view it as an intrusion into their home. This type of technology could possibly be improved by using the information gathered to communicate with the user in real time, helping to promote healthier daily behaviour.

Another type of activity monitoring is a wandering detection system. Wandering is the term used to describe a person with a cognitive impairment (such as dementia) leaving the safety of their home during an episode of confusion or without adequate preparation. There are two types of wandering detection system: one to raise an alert when a person wanders off and one to help locate them again. There are a wide variety of approaches to wandering detection, almost all of which rely on sensors in the home. Although there is a great deal of variety in the technology and a wide range of manufacturers, almost none of the technology attempts to communicate with the wanderer.[11] There are few items which actively try to prevent wandering, and those items generally work through a remote secure entry and exit system [77]. Secure entry and exit can also be used to protect occupants from bogus callers.

Of the devices which help to locate a wanderer, most are GPS devices housed in a plain plastic keyring. Of the 14 devices discussed on the Living Made Easy website, 2 are housed in a mobile phone, 1 in a bracelet and 1 in a watch; the remainder are simple plastic boxes. Many of the devices also incorporate a panic button to give the user and carers piece of mind. However, it is easy to image that someone with a memory-related cognitive impairment would lose the device or forget about the panic button when they are lost or confused. This makes the technology both less appropriate and effective, and highlights a key area for improvement in relation to Thesis Question 4.

Other types of activity management system can be built on top of telecare or telehealth systems by including additional sensors. Tunstall’s offerings in particular are designed

[10] Just Checking Ltd, `http://www.justchecking.co.uk`.

[11] That this may sometimes be the best course of action, as a sudden disembodied voice could cause unnecessary distress.

to be highly modular so that the system can be tailored to the needs of the care recipient [200].

2.2.5 Environment Management

> *"Environmental Control Systems can enable you to operate everyday domestic appliances and mechanisms by remote control."*
>
> — Living Made Easy [131]

Environment management systems are designed to help care recipients to manage their home, or to provide carers with information about the home environment so they can step in if needed. Examples include door, window and temperature sensors for observation, but also includes technology allowing residents and carers to remotely control secure entry systems, home appliances, blinds, curtains and more. For example, technology that follows the X10 standard can be used to control most simple household electrical appliances.

The Tunstall catalogue [200] includes several devices which can be used to monitor the home, but they do not produce a 'universal remote' style device for appliance control. However, a wide range of such devices are available from a number of other manufacturers. These devices cover a wide range of interaction styles. Some use simple oversized buttons while others run on tablet computers or smart phones. Many of the devices provide visual and audio feedback along with the tactile feedback involved in pressing a physical button. In general these devices provide an excellent way for users to remotely manage appliances in their homes. These devices also show a lot more thought in terms of aesthetics. The Amdi Tech-Talk Plus Environmental[12], shown in Figure 2.3a has programmable speech messages, interchangeable overlays and a detachable grid. Another interesting device is the QED tilt switch[13] shown in Figure 2.3b. This device is worn on the head and appliances are operated by the user tilting their head in certain directions.

These devices demonstrate an acute understanding of the unique needs of the user, taking into account sensory, cognitive or physical impairments to make them both appropriate and effective. Unfortunately, many of these devices are surprisingly expensive; the Living Made Easy website shows the most expensive device is £895 and the cheapest

[12] Amdi USA, `http://www.amdi.net/`

[13] QED tilt switch, `http://www.qedonline.co.uk`.

(a) Amdi Tech-Talk Plus Environmental. **(b)** QED Tilt Switch.

Figure 2.3: Two environmental control devices.

£24[14]. As those prices are only for the interaction device, the full home management system would cost considerably more, and would require the aid of an engineer to install.

The high prices and complexity are generally due to the high level of customisation on offer. It is feasible that technology equipped with multiple modalities and designed to be flexible could be made much more affordable than such bespoke technology. The inclusion of multiple modalities would provide several options for interaction which would only require configuration, so a high level of customisation and effectiveness could in theory still be reached by such technology.

A slightly different example of an environmental management system is the i-Cue [77] system developed by Halliday James. i-Cue is a modular system that is able to effectively manage a wide range of things in the home including the heating system, taps and electrical appliances. It can also monitor activity and interact with the resident, making it a powerful solution applicable to a variety of situations. The i-Cue system is more in alignment with the thesis questions set out in Section 1.2, although it does not focus on using multiple methods of interaction as a core part of the technology.

[14]Prices retrieved on the 31st January 2013. Source: Living Made Easy [131].

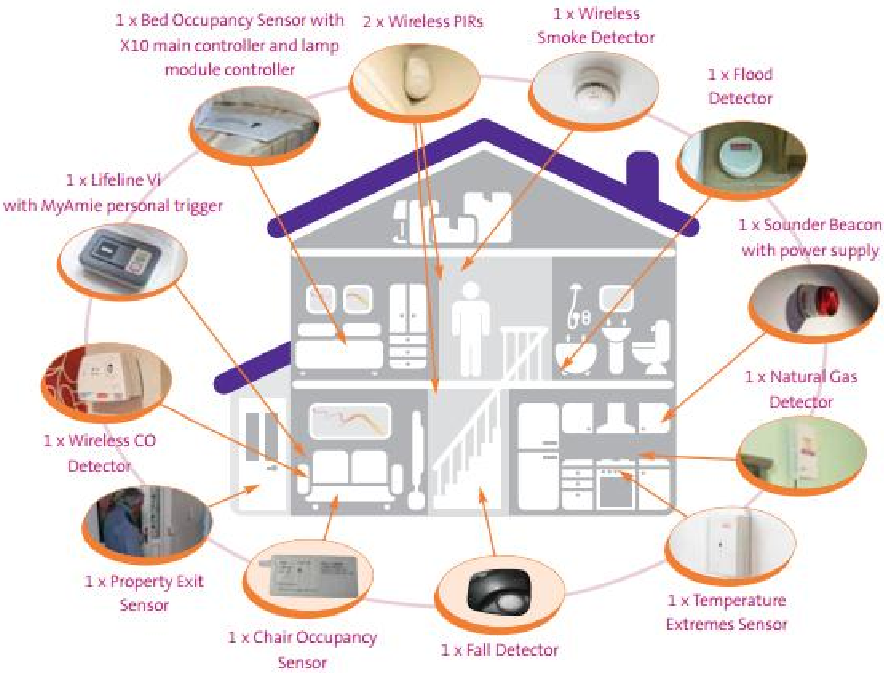


Figure 2.4: A telecare solution suggested by Tunstall in their catalogue [200, p. 44].

2.2.6 Observations on the Home Care Industry

This section presented some of the technology produced by the home care industry, allowing several observations to be made. Firstly, very little emphasis is placed on the resident of the house: most of the technology provides assistance by calling in another person. Few devices actively interacted with the care recipient to provide care. Thesis question 4 asks how this technology could be made more appropriate and effective, and including the user in their care is likely to have a positive impact on both factors. Another way to make the technology more appropriate would be to respect the user's preferences and the environment of the home. If home care technology is designed to communicate with the user more often, users should be able to configure those interactions to suit their preferences, unique needs and environment.

However, not all devices avoid interacting with the user. One notable exception is the AbleLink system sold by Halliday James[15]. This is a software solution that runs on a mobile phone, but which uses visual and audio reminders to help the user organise their daily activities. Reminder solutions, which aim to help promote a higher standard of living by encouraging self-reliance, are actually surprisingly rare. Tunstall offer a device called the LifeLine Vi Plus [200], which is able to provide custom speech reminders via bespoke programming.

[15] Halliday James, http://www.hallidayjames.com.

Another observation that can be made is that almost every system requires the aid of an engineer to install and configure. The systems are too complex for carers and family to effectively manage, which limits the flexibility of most of these solutions despite their modular design. As care needs can be expected to change with time, a reliance on trained personnel poses a problem. Thesis question 3 asks how home reminder technology can be designed to best utilise multiple types of interaction. Existing work shows that modularity alone is not sufficient: the role of the engineer must be reduced by creating technology that is more flexible and robust. More dynamic technology equipped with multiple forms of interaction could also provide the option of end-user configuration for simple changes, which would further improve the appropriateness of the technology.

Figure 2.4 shows a telecare solution suggested by Tunstall based on the products available in their catalogue [200]. This suggested solution supports many of the conclusions drawn here: there are a large number of white plastic devices that are unlikely to compliment the home's aesthetics; the devices are distributed throughout the house in a variety of ways, which would most likely require an engineer to install and configure; and the suggested solution includes no components (other than a flashing alarm in the kitchen) that communicate with the resident.[16] With respect to Thesis Question 4, it is clear that there is a lot of space for improving the effectiveness and appropriateness of current home care technology.

2.3 Home Care in Research

Section 2.2 showed that industry products are built around several different models, with the most popular model being a modular network of sensors that gathers lifestyle data for stakeholders and raises the alarm in emergency situations. As discussed in Section 2.2.6 this approach has several shortcomings, most notable of which is that the technology rarely interacts with the resident.

While there is considerable research work that aims to tackle this, it generally does so by working towards a 'home of the future' (or smart home) concept, where the overt goal is increasing the comfort of the resident. This work is described in Section

[16] The LifeLine Vi device in the diagram is the standard model, which does not have the ability to deliver reminders like the 'plus' model.

2.3.1, followed by Section 2.3.2 which considers alternative approaches. Section 2.3.3 summarises and explores the themes observed in home care research projects.

2.3.1 Smart Home Projects

The 'home of the future' concept, also known as smart homes, generally describe environments rich with sensors and actuators tasked at providing a high level of domestic comfort.

> *"With sensor arrays and digital displays embedded into most surfaces, the home begins to discover their patterns of activity and tries to anticipate what they might need or want. At first, it gets it only about half right, but within several weeks it begins to fit like a glove. It adjusts the ambient light for reading a book in the afternoon, keeps tuna fish on hand in the pantry, monitors their nutrition, and suggests new films that they may enjoy."*
>
> — The Home of the Future [122, p. 2]

Many researchers have suggested that similar smart environments could be used to provide home care. While current home care products outfit the home with technology, they do not attempt to create a smart home of the type described by Larson [122]. Thesis Question 4 asks how to make home care technology more appropriate; is the home of the future paradigm the way to solve this problem? This section presents a review of several smart home research projects concerned with care in order to explore this question.

The Gator Tech Smart House

The Gator Tech Smart House[17] at the University of Florida is a prime example of the 'home of the future' concept. Helal *et al.* [85] stated several aims for the Gator Tech Smart House project:

- "*[to create] a programmable space specifically for the elderly and disabled*" [85, p. 64].
- "*create assistive environments that can ... enact mappings between the physical world and the remote monitoring and intervention services*" [85, p. 64].

[17] The Gator Tech Smart House, http://www.icta.ufl.edu/gt.htm.

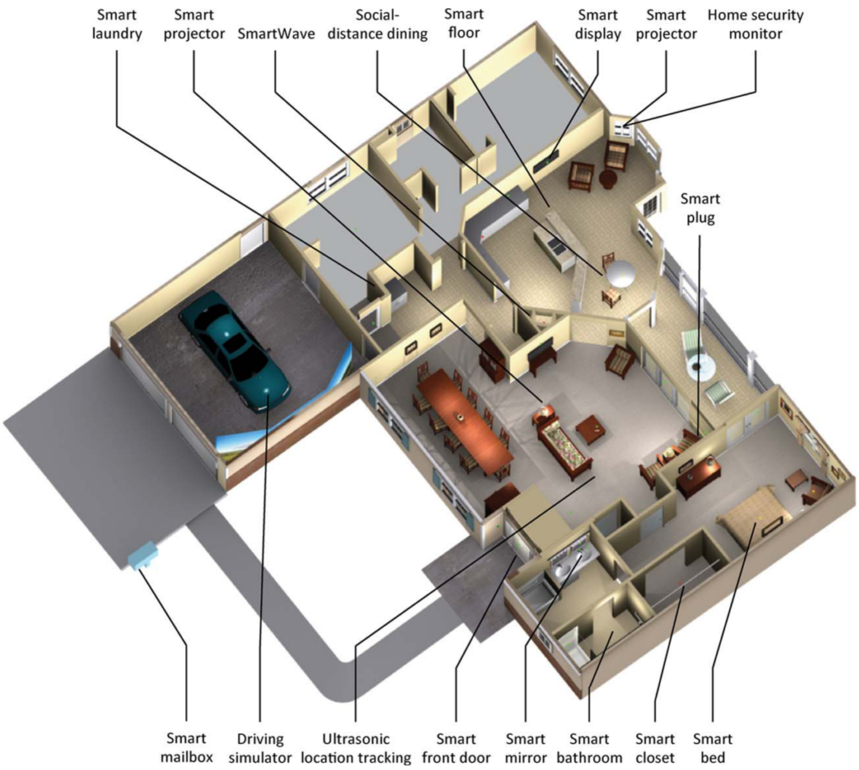


Figure 2.5: An illustration of the Gator Tech Smart House showing where smart technology is located.

- "*to outlive existing technologies and be open for new applications that researchers might develop in the future*" [85, p. 73].

- "*... our goal is to create a 'smart house in a box': off-the-shelf assistive technology for the home that the average user can buy, install, and monitor without the aid of engineers*" [85, p. 73].

These aims appear to be in alignment with the requirements set out in Section 2.1; however, the implementation of the Gator Tech Smart Home fell short of many of these aims. This is partly due to the technology used to implement the Gator Tech Smart House. Shown in Figure 2.5, smart technology inside the house includes a smart mailbox[18], smart front door [19], smart bathroom[20], smart closet[21], smart bed[22], smart

[18] "*The mailbox senses mail arrival and notifies the occupant.*"

[19] "*The front door includes [an RFID] tag for keyless entry by residents and authorized personnel ...*"

[20] "*[The bathroom] includes a toilet paper sensor, a flush detector, a shower that ... prevents scalding, and a soap dispenser that monitors occupant cleanliness and notifies the service center when a refill is required ...*"

[21] "*[The closet will] make clothing suggestions based on outdoor weather conditions.*"

[22] "*[The bed] has special equipment to monitor the occupants' sleep patterns...*"

blinds[23] and more. While the "*Smartwave*" microwave [178] is automatically able to cook microwave meals, which is likely to be very useful, it's not clear what benefit the driving simulator[24] in the garage would provide to a home care user.

Helal *et al.* [85] argue that one of the primary contributions of the Gator Tech Smart House is the technology used to create a smart environment from a wide range of heterogeneous technologies, including the necessary middleware systems to manage sensors and actuators. However, the original implementation of this was later found to fall short of the required levels of robustness and portability [82].

Helal & Chen [82] outlined 3 important lessons learned from the initial Gator Tech Smart Home implementation: *(1)* that it was difficult to make the heterogeneous technology work together due to hardware and software differences, *(2)* that the hard-wiring of technology made installing and managing the system labour-intensive, and *(3)* that due to these two issues the system was unable to evolve with changing requirements. The Atlas middleware layer [121] was developed to address these shortcomings, which used lightweight connectors and OSGI to create hardware and software 'packages' which would let the technology be more flexible. This would make one of the stated aims of the project more feasible; that users should be able to buy, install and configure the technology without an engineer.

Helal *et al.* [83] asserted that a live-in trial demonstrated the value of smart technology, in particular for reducing isolation in older people. However, this trial was carried out with a single unimpaired user over one week, making its findings unreliable. Helal *et al.* also argued for further work on how such smart home devices can be configured and managed by end users.

Another aspect of the Gator Tech Smart Home project was persuasive technology to improve lifestyle habits. Lee *et al.* [125] argued against the current telehealth trend for sensors that transmit data to health professionals, and suggested that more communication between users and technology could be beneficial. In particular it is suggested that 'persuasive' technology could be used to improve the general well-being of telehealth technology users, particularly elderly users. One suggestion is the Action-Based Behaviour Model [126], which involves communicating information about conditions or general health and well-being to the user along with the formulation

[23] "*[The blinds] can be preset or adjusted via remote device...*"
[24] "*The garage has a driving simulator to evaluate elderly driving abilities and gather data for research purposes.*"

of goals. Lee *et al.* hoped that the combination of prompts, reminders and lifestyle observation data would persuade users into carrying out more ‘positive’ activities.

Another suggestion is that domestic robots might be an excellent way to influence user behaviour due to their “natural human-like communications” [125, p. 6]. Lee *et al.* [127] carried out a survey of suitable domestic robot technology, comparing their features. The robots were designed primarily around communication, and are unable to affect the physical environment in any meaningful way. Depending on the needs of the user, a robotic reminder system might be effective while also helping to alleviate any feelings of isolation or loneliness. Unfortunately, of the eight robots evaluated only 2 were commercially available and supported the English language. More work is needed to fully understand the role that robots might play in a home care system, and that work is outside the scope of this thesis.

Practicality is an issue with the Gator Tech Smart House. Much of the technology would be destructive to install, such as the “smart floor” and “smart sockets”. The floor itself is surprisingly inexpensive[25], however the authors suggest that “*We do not have to factor the price of the raised floor, which is comparable to other types of residential flooring, into our cost analysis because it is a fundamental part of the Smart House...*” [85, p.72]. The authors later acknowledge that the installation of these materials was labour-intensive and time-consuming [125], which suggests that even with the Atlas technology older users might be highly inconvenienced by the installation.

Thesis Question 4 considers how to make home care technology more appropriate and effective. The Gator Tech Smart Home is technologically advanced, yet there is little evidence that the technology within would provide a significant benefit to home care users. Some parts of the technology are likely to be inappropriate for use in the home, such as the smart floor, due to the amount of work that would be needed to install it in the home. Although one of the stated aims was to create technology that can be bought ‘off the shelf’ and installed without engineers, it is clear that much of the technology in the Gator Tech Smart Home did not meet that requirement.

The MIT House_n Consortium

Intille *et al.* [99] argue that the MIT House_n project[26] avoids the traditional smart home paradigm by focussing on “communication, not automation”. One of the projects

[25] $4 per square foot, or $43.06 per m^2: Helal *et al.* [85, p. 72].
[26] The MIT House_n Consortium, `http://architecture.mit.edu/house_n/`.

Figure 2.6: Photographs of the MIT PlaceLab from Intille *et al.* [102, p. 1942]. Each of the black squares shown in the pictures covers sensor circuitry. While the sensors themselves are not intrusive, the installation process would be disruptive.

undertaken as part of the House_n project was the development of lightweight 'tape-on' sensors that could be used for ethnographic studies in the home.

Beaudin *et al.* [18] carried out a field trial with these sensors, but found that invasion of privacy was a recurring theme in feedback. They also found that "there is no normal week"; despite claims from participants that their home routines would be uneventful, the authors found that home routines were highly malleable and would change based on factors such as "job loss, diet change after diagnosis of an allergy, holidays, illness, guests, pet illness, changes in the weather, a sudden business trip, taxes coming due, and preparations for a new roommate" [18, p. 1361].

The MIT House_n project also created a live-in laboratory called the PlaceLab. Intille *et al.* [102] describe the PlaceLab as a reconfigurable space with several sensor networks used to record everything about the inhabitants. The PlaceLab also interacts with several mobile sensor systems including wearable sensors. As shown in Figure 2.6, this technology would be well-hidden in the home. However, it is also clear that such sensor technology would require a destructive installation in an existing home, primarily due to the position and number of the sensors.

Morris *et al.* [157] suggested using this sort of sensor network to provide early detection of health problems by monitoring regular routines and behavioural properties, then watching for unusual deviations. This is what the 'Just in Time Carer' discussed in Section 2.2.4 attempts, although it relies on a human element to interpret any deviations. Fully automating this process is likely to be extremely difficult due to the natural plasticity of domestic routines, as Beaudin *et al.* [18] found in their sensor field trial.

Beaudin *et al.* [17] conducted a series of interviews regarding this sort of longitudinal monitoring. It was found that the technology was more likely to be accepted by users if it was provided as a customisable tool for self-improvement than as a health-monitoring or medical communication device, much like Lee *et al.* [126] suggested with their Action-Based Behaviour Model. Considering Thesis Question 4, this provides additional evidence that more appropriate technology includes and interacts with the user.

Intille *et al.* [99, 100] suggested that the best way for this technology to support older users would be "just-in-time prompts": messages delivered by a context-aware technology to provide some level of care. Intille *et al.* [100] outlined some of the challenges that such technology would face, including four requirements for the prompts themselves: they must be simple and easy to understand, delivered at the right time, in the appropriate place and in a non-irritating way. Intille *et al.* also argued that they should be personalised to the user to maximise their effectiveness.

Nawyn *et al.* [163] evaluated the ability of "just-in-time prompts" and a dynamic context-aware interface on a universal remote control at reducing sedentary behaviour in the home. They did this by providing prompts to promote 'wanted behaviour' and attempting to prevent 'unwanted behaviour' by both disrupting it when it occurs and encouraging incompatible activities (*e.g.* encouraging a user to do chores or play games to stop them from watching TV). They also attempted to encourage self-monitoring of time and activities. Their prototype was evaluated in the PlaceLab with mixed results; the general conclusion from the work was that these methods *could* enact a change in behaviour, but poor design and a poor understanding of the user would be likely to result in rejection. More flexible technology, in particular technology that included multiple modalities, would provide several ways to ensure that a system aligns the users needs and stays aligned over time.

The Aware Home Research Initiative

The Aware Home Research Initiative[27] is a project at the Georgia Institute of Technology. As part of the project a live-in laboratory was created called the Aware Home [118]. The Aware Home has two independent identical living spaces with a shared basement area. One of the stated aims of the project is to "*. . . to design a system that provides a type of monitoring currently supported by an assisted living center for those individuals that do not demand frequent medical help or services that could only be provided by another person*" [118, p. 196]. It is also suggested that the overall aim of this technology is to be installed into the existing homes of care recipients.

Mynatt *et al.* [159] described how the sensors in the house would be able to detect dangerous situations using a mixture of dedicated sensors. This paper was written in 2000, and industry has effectively caught up with companies like Tunstall providing this sort of solution as discussed in Section 2.2.1. However, there is a notable difference between this research technology and that available commercially: the research work aims to create technology that will fit into the home without disrupting domestic routines.

A good example of this is the Digital Family Portrait technology created by Mynatt *et al.* [161]. This uses sensors in the home to gather lifestyle data about the resident, much in the same way that the 'Just in Time Carer' described in Section 2.2.4 does. While the 'Just in Time Carer' uses a web interface to display information, the Digital Family Portrait displays the information graphically. A picture of the resident is shown in the centre of the digital frame while the frame itself fills with icons as activities are recorded. The frames can be fully customised to suit the person, which is in stark contrast to much of the technology provided by industry which offers very little customisation. However, much like a great deal of the technology commercially available, the Digital Family Portrait is designed to give peace of mind to carers, and not the resident.

Much like Lee *et al.* [125], Rogers & Mynatt [176] also advocated for technology that promotes communication. Unlike Lee *et al.* however, Rogers & Mynatt do not refer to communication with the technology; they instead advocate technology that increases communication between care recipients and carers, and give the Digital Family Portrait as an example of this. The Peek-a-Drawer [186], which does promote two-way communication, is a self-contained unit which seems to have little benefit as a home care device or smart home component. Another project called Cook's Collage [198, 197]

[27] The Aware Home Research Initiative, http://awarehome.imtc.gatech.edu/.

aims to help resume cooking tasks after an interruption by providing visual cues to indicate the previous activity; while it is conceivable that this might be useful in certain home care scenarios, this is not discussed in these papers.

Mynatt *et al.* [160] argued for technology which considers the needs of older users, in particular, technology to support communication, memory aid (reminder prompts) and mobility. In this paper Mynatt *et al.* presented three pieces of technology to older people; the Digital Family Portrait, Cook's Collage and a gesture-pendant. When asked for feedback on the devices, privacy was revealed to be a significant concern, similar to the findings of Beaudin *et al.* [18]. The participants also expressed doubts about the value that this technology would bring them, with the Cook's Collage seeming to garner the most interest. Oddly, the authors avoid discussing that the Cook's Collage was the only piece of technology that provided any direct benefit to the resident.

Kientz *et al.* [119] summarises the projects that were carried out as part of the Aware Home Research Initiative, describing 13 different projects. Of those projects, only 2 are clearly designed to help older users live independently; a smart mirror that helps to track items (such as medicine bottles) and the 'Technology Coach', designed to provide medical device training [158]. While the project's researchers frequently argue for technology that supports the user themselves, the work carried out as part of the project does not appear to be firmly grounded in the needs of a home care recipient. Like the other smart home projects, practicality is an issue: the Aware Home was designed to provide a flexible and reconfigurable environment for experimentation, with easy access to conduits and sensors. This sort of access is unlikely to be available in the homes of most care recipients. However, technology has moved on considerably since the Aware Home was constructed with wireless sensors more readily available than they were in 2000.

The Millennium Home Project

Dowall & Perry [55] described the Millennium Home project, which was intended to provide support for older users living at home. The given aims of the project were to create smart home technology much like that in the Aware Home and PlaceLab, but with a greater focus on affordability, flexibility and practicality. Perry *et al.* [171] criticised the lack of novel research into technology for supporting elderly users. In particular, they note that most of the exiting work in this area represents "*add-ons to generic 'home of the future' automation systems, rather than systems primarily designed to support independent living for older people*" [171, p. 258]. Another goal for the project was that the technology should not require the user to carry anything, *e.g.* a pendant or mobile device.

Perry *et al.* [171] argued that users of this technology are likely to have very diverse needs and abilities, and as such interaction methods should be available that can accommodate the unique needs of the user. Adaptability was also promoted as an important factor in a home care system, allowing the user to modify the system quickly with changing social or personal needs without preventing it from working. Perry *et al.* suggest that multimodal interaction is key to providing adaptable and robust dynamic technology. Although Perry *et al.* limited their system design to "*support a cognitively fit and able-bodied user*" Perry *et al.* [171, p. 262], they suggest that improved multimodal interaction techniques could allow the Millennium Home system to support users with impairments.

Observations on Smart Home Projects

Thesis Question 3 considers how home care technology can be designed, and it's clear that many of the smart home research projects understand the need for flexible technology in this regard. There's also a much stronger emphasis on customisation and on making the technology appropriate for the home, demonstrating some of the ways that home care technology could be made more appropriate. Several of the researchers advocated for technology that communicates with the resident to help improve their quality of life, which was shown to be highly appropriate.

Adaptability is given as an important requirement in all of the projects described. The reasons that this is important include the need to evolve technology over time and to customise the technology for specific user needs. Most of the projects did not carry out evaluations with older users despite creating live-in laboratories.

Practicality is also a concern, as with the exception of the the Millennium Home project, it is unclear how the smart technology could be installed into the home in a cost-effective and non-destructive manner. In addition, there is little evidence that this expensive technology will actually be able to provide a high standard of care, as much of the technology discussed has not been tested in a care scenario.

The work of Perry *et al.* [171] seems the most realistic; it demonstrates a strong understanding of the needs of the user, the design problems involved and the general goals of care technology. It also deeply rooted in the practicalities of smart home technology, particularly with respect to the cost involved in buying, installing and configuring such systems. Perry *et al.*'s work will be key to addressing Thesis Question 3, which considers how home care technology should be designed to use multiple forms of interaction.

2.3.2 Alternative Approaches

There are several alternative approaches to providing care at home that do not follow the smart home paradigm. This section will explore some of these alternatives.

Smart Residential Care

Stanford [190] suggests that instead of creating technology for heterogeneous home environments, we could instead focus on building residential care facilities designed around smart technology. Stanford refers to an existing home that allows carers to track residents and remotely monitor their weight, bathroom activities and sleep patterns. Stanford suggests that after the initial move, elderly residents can continue life 'as normal as possible' in their new environment. Most research in this field aims to keep older people at home, but for those with care scenarios too complex for a care system to deal with, residential care might be necessary. Smart technology in this situation is likely to provide a benefit to both the standard of living in residential care and the economic issues raised in care scenarios. However, for the majority of users moving into residential care is not likely to be the most appropriate course of action, as highlighted by the My Home Life report [194].

Mobile & Wearable Technology

Is complex technology really necessary for improving home care? Research by Bickmore *et al.* [22, 23] suggests it is not. The authors developed a simple mobile application that reminds users about simple health-related tasks, such as taking medicine or getting some light exercise. Reminder devices such as this could conceivably help an elderly user manage their lifestyle, offering them a greater degree of independence at home. CybreMinder by Dey & Abowd [54] is another example of a mobile-phone based reminder system.

In terms of thesis questions 3 and 4, it's likely that small reminder systems like this will be much more appropriate for the majority of users than an expensive and complex smart home solution. There are also several examples of mobile phone technology being used in conjunction with smart home sensors, or even as a sensor itself, *e.g.* Jorge [111] and the mPCA project [84, 72].

Mobile phones are a useful tool when considering how to provide home care, as they generally provide at least three ways to interact (visual, aural and tactile), they are inexpensive, highly programmable and have a built-in ability to communicate with external parties. Yet there are few telecare or telehealth devices based around the mobile phone, with the exception of the Halliday James AbleLink discussed in Section 2.2.6. This is because mobile phones have not been designed with care provision in mind; as such they have several limitations, *e.g.* depending on the care recipient to charge the phone.

A common theme for telehealth research is the use of wearable technology. The AWARE architecture [132] is one example of a wearable system that aims to provide monitoring of health metrics that can help users and carers manage conditions. Another project called SAPHIRE [81] is similar but is aimed at promoting remote observation of of conditions that would normally keep a person in hospital under observation, such as recovery after operations. These systems generally consist of wearable sensors and some sort of central processing device such as a mobile phone or PDA. Due to the reliability demands of these scenarios, in particular for SAPHIRE, a great deal of work has been carried out to ensure that such systems are as robust as possible through the implementation of the self-managed cell model [116]. Wearable technology could certainly also have applications in elder care, especially if woven into clothing; however it has been suggested that dedicated wearable devices might be difficult to deploy in a home care setting [171].

2.3.3 Observations on Home Care Research

Thesis Question 3 considers how home care technology can be designed to best utilise multiple interaction methods. In both industry and research, systems made from modular components have been central to home care designs. This is because the home environment and user's needs will change over time, as noted in Section 2.1. Research projects have made several attempts to increase the flexibility of these modular designs, yet many of them have fallen short [125]. That may be because many research projects focus on the smart home paradigm, while it is likely that far less complex technology would be more appropriate for providing care at home [22, 23, 54, 171].

Thesis Question 4 asks how home care technology can be made more effective and appropriate. A recurring theme in research work is that of reminder technologies: simple systems that provide notifications to help care recipients carry out their daily activities [17, 22, 54, 100, 125, 157, 163, 171]. Reminder technologies are likely to be a very appropriate way to provide care for older users, as they include the user and promote independence and healthy daily routines. This type of care technology is likely to result in a much higher standard of living than technology which simply calls for outside help in an emergency situation. Reminder systems will be explored in more detail in the following section.

2.4 Notifications & Reminders

A large number of researchers have advocated for reminders, persuasive technology or just-in-time prompts[28] to promote healthy ageing at home for older adults [23, 54, 100, 125, 138, 146, 160, 171]. Notifications and reminders have been shown to be effective in a number of situations. McFarlane & Latorella [141] pointed out the importance of notifications in the workplace, in particular with respect to monitoring information and negotiating time between activities. Research into interruptions in the workplace and car are plentiful, and while there is less work that uses older people or the home, there are many interesting examples. Several smart sinks have been shown to be effective at helping people with severe dementia complete a hand-washing task [6, 27, 150]. Tran *et al.* [196] claimed that their Cook's Collage technology helped older users with cooking by helping to manage time and 'position' in a recipe. Bickmore *et al.* [22] demonstrated that mobile device reminders would help with health regimen compliance. Nawyn *et al.* [163] demonstrated the effectiveness of a mobile notification device in breaking sedentary activity habits in a live-in laboratory.

Such projects show that there is great potential for reminder technology to help improve independence and quality of life. More work is needed however to fully understand how reminders could be applied to promote healthy and independent ageing in older people. This section will explore reminder technology, starting with an examination of the psychology of notifications and the positive and negative effects they might have in Section 2.4.1. Following this, Section 2.4.2 will explore the design issues surrounding

[28] Please note that for the remainder of this thesis, the terms 'notification' and 'reminder' will be used interchangeably to refer to any short message sent to the user for the purposes of prompting a response or providing information. For more information please see the glossary on page 358.

reminder systems such as how and when to deliver such reminders along with issues surrounding the configuration of the technology. Finally, Section 2.4.5 summarises the observations made here.

2.4.1 The Psychology of Notifications

Notifications are extremely useful; when multitasking, they can be used to manage attention, time and information to support high performance in multiple tasks. Research has shown that in certain contexts notifications can lead to improved speed [39, 115, 134, 180] when carrying out multiple tasks. Speier *et al.* [189] showed that notification can improve decision-making for simple tasks, although also noted a deleterious effect for more complex activities. The act of interrupting a task has a number of interesting psychological effects which will be explored in this section.

Research into interruptions is a rich field that pre-dates the modern computer by many years. As such, there is a great deal of literature to provide an insight into the risks and benefits of interruptions. The first major piece of research into the effects of interruptions was carried out in 1927 by Zeigarnik [230] after observing waiters in a café. Zeigarnik theorised that an incomplete task would remain 'in tension', and that people are driven by a need to finish what they have started (this is called the 'tendency to complete'). This desire causes people to retain more information about an interrupted task than an uninterrupted task. This would come to be known as the 'Zeigarnik Effect'. Despite decades of further research and investigation, the Zeigarnik effect remains relevant today and Zeigarnik's seminal paper "On Finished and Unfinished Tasks" [230] is frequently cited in modern interruption literature.

Speier *et al.* [189] discovered that interruptions can sometimes have a positive performance impact, depending on context. Specifically, Speier *et al.* found that for simple tasks interruptions would improve decision making performance. The authors theorized that in cognitively non-demanding tasks, the subjects would 'distract themselves' by thinking about other things. They believed that interruptions pushed the cognitive workload of a simple task above the threshold required for full attention, resulting in improved performance. It was found that this only held true for simple tasks, and that interruptions on a more complex task had a detrimental effect. They also found that a higher frequency of interruptions had a significant negative effect on performance.

Kapitska and Blinnikova [115] reported the results of an experiment where they observed that participants would try to ensure the same time is spent on a task regardless of whether they were interrupted. They did notice an increase in error rate, suggesting that participants were willing to sacrifice accuracy and performance in order to 'stay on top' of their workload.

Burmistrov & Leonova [39] found that interruptions actually led to an increase in the speed at which the main task was executed. This has also been observed by other researchers, such as Mark *et al.* [134], who also noted an increase in stress levels. A similar effect was observed by Sanders & Baron [180], but with distractions as opposed to interruptions.

Cohen [46] explored the physical impact of interruptions, specifically relating to stress. After reviewing other literature in the field, Cohen identified that interruptions could create stress, the effects of which would linger long after the interruptions have been dealt with. Cohen was able to show that making interruptions more predictable or controllable could reduce this unwanted stress, a finding which was confirmed by Carton & Aiello [42].

Similarly, an experiment by Bailey, Konstan & Carlis [12] found that the presence of reminders increased annoyance, anxiety, and perception of the difficulty of the given tasks. They also noted that these negative effects can still be felt long after the task has been completed.

Brehemer *et al.* [29] found that older users people were disrupted more by interruptions than younger people were. Specifically, an older persons' performance on a cognitively demanding task decreased much more when interrupted than younger people were. This could mean that in a home care scenario the effects of an interruption could potentially put the user at risk, *e.g.* by creating an opportunity for mistakes to be made when preparing food or cleaning.

The existing research suggests that there is a wide range of negative effects that could be produced by notifications, such as increased speed, more errors and higher stress. The research also suggests that these effects could be more pronounced in older people. Berg *et al.* [21] examined the causes of falls in the homes of older people and found that most were avoidable and caused by haste and distraction. It is important that any reminder technology placed into the home of an older person does not put them at risk by inducing changes to their behaviour. Preventing such a situation relies on a good understanding of the psychological effects of notifications and task switching.

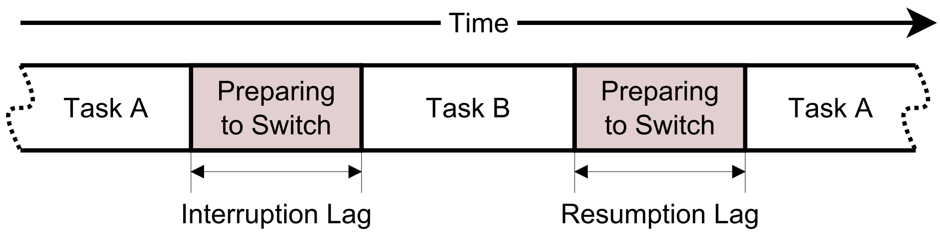


Figure 2.7: Diagram showing the process of task switching. The time taken to prepare to change task is called the interruption lag and the time taken to resume the original task is called the resumption lag [45, 195].

A Mental Model of Task Switching

Cellier & Eyrolle [43] investigated how people coped with managing multiple tasks, and found that two tasks would interfere with each other. In particular, they found that interrupting a task to carry out another required 'saving' the state of the current task and the reallocation of resources. Any disruption to this process led to an increase in processing time and the number of mistakes made.

When task switching takes place after a notification, there is a short delay between the notification switching to the secondary task. This pause is called the interruption lag [45, 195], and this time is used to take a 'mental snapshot' of the task being stopped. After the secondary task has been carried out, this mental snapshot is 'reloaded'; the time taken to do this is called the resumption lag [45, 195]. This process is shown in Figure 2.7.

Trafton *et al.* [195] found that the longer the interruption lag, the shorter the resumption lag; *i.e.* that when given more time to prepare, task switching becomes easier. Clifford & Altmann [45] found that providing cues during the interruption lag to facilitate 'note taking' will reduce resumption lag for a suitably complex task, but that asking participants to take physical notes would increase the resumption lag.

Gillie & Broadbent [71] performed a number of experiments to explore how people resumed an interrupted task. Their experiments showed that the ability to resume a task did not appear to be related to either the memory load at the time of interruption or the length of the interruption itself. Instead, they found that the nature of the secondary task had the largest impact on task resumption. They claimed that highly complex secondary task or a secondary task similar in nature to the primary task would have the greatest disruptive effect. That is, the 'mental snapshot' prepared during the interruption lag will degrade if the secondary task is either too similar or too complex. Scott *et al.* [182] showed that the resumption lag can very between 20 and 60 seconds based on the complexity of the interruption and the type of assistance offered during task resumption. However, those times are for a complex organisational task: simple activities are likely to be resumed much quicker.

Hodgetts & Jones [87, 88, 89] found that resumption lag increased along with the length of the interruption. This effect is called the goal activation decay, and was first proposed by Altmann & Trafton [2, 4]. Monk *et al.* [155] carried out an experiment which confirmed that resumption lag increased with interruptions of any length, finding that interruptions as short as 0.25 of a second were just as disruptive as interruption 5 seconds long. A later experiment by Monk [153] showed that increasing the frequency of interruptions reduced the resumption lag when task switching.

In summary, when a person switches from one task to another they 'save' their mental state prior to the interruption and try to restore it afterwards. This process is usually very quick, generally taking only a few seconds. Yet if not enough time is spent preparing to switch task, then resuming the task later becomes much more difficult. However, even if sufficient time is spent preparing, the 'mental snapshot' will degrade if the secondary tasks are mentally demanding or are highly similar to the interrupted task. Home care reminders should strive to ensure that this process can be carried out in full, helping to ensure minimal disruption to household activities.

Issues in Multitasking

Miyata & Norman [152] noted that there were many external and psychological factors involved when multitasking. Specifically, the most important issues were:

Memory Demand

There are many different types of memory[29], but Miyata & Norman [152] point out the most important factor when considering multitasking is that short term memory is limited. A good example of this is the well-known 'magic number' of 7 ± 2, which was suggested by Miller [151] in 1955 as the amount of information the average person can hold in their working memory at any given time.

[29] There are also several schools of thought concerning the type and functionality of memory, but that is outside of the scope of this literature review.

Planning Opportunities

Miyata & Norman [152] suggest that as most primary activities can only be carried out sequentially, even when multitasking, planning ahead helps to organise resources and time. Research by Cutrell *et al.* [51] found that the stage of the primary task in which the interruption is delivered had a significant impact on how disruptive the interruption is. In particular, they found that interruptions during the early planning stages of a task were the most disruptive. Carton & Aiello [42] found that the interruptions which could be predicted resulted in better performance.

Experience with Task

Miyata & Norman claim that there are two types of task: Conscious tasks (*e.g.* writing) are those which require full attention and suffer significantly when interrupted, while subconscious tasks (*e.g.* walking) are said to require little to no mental resources and are not massively disrupted by an interruption. Miyata & Norman [152] suggested that a large number of subconscious tasks can be carried out simultaneously, limited only by what is physically possible: *e.g.* walking and talking are two separate activities that do not interfere with each other. Miyata & Norman also suggested that with diligent practice some conscious tasks will become subconscious tasks, such as playing certain musical instruments. Cades *et al.* [40] discovered that with practice and/or exposure, the disruptive impact of interruptions will decrease, resulting in quicker task switching and fewer errors.

Attentional Focus

Miyata & Norman [152] suggest that there are two different types of task: task-driven and interrupt driven. Task-driven tasks require focussed attention and tend to be more heavily disrupted by an interruption, *e.g.* reading. Due to the attentional demands of the task however, it is harder for external information to actually cause an interruption. Interrupt-driven tasks are those that are supported by large quantities of external information, *e.g.* driving. Such tasks are more resilient when faced with an interruption due to their dynamic nature; however as these tasks involve being receptive for external stimuli, it is much easier for unwanted information to interrupt the task.

Combining Multiple Tasks

Miyata & Norman [152] suggest that the type of activities being carried out at the same time will have a significant impact on multitasking ability. While Miyata & Norman only cover this superficially, other research has shown that this is the case; Gillie & Broadbent [71] found that the similarity of two tasks and their complexity would have a considerable impact on a person's ability to carry out the two tasks in parallel.

Reminders

In this context, Miyata & Norman [152] define reminders in two ways; firstly, it is some sort of automatic event to help manage time and secondly it is the use of some external markers to serve as an aid to task resumption. Miyata & Norman point out that reminders can be used to overcome the limitations of short term memory. Research has shown that people use external cues to help them resume an interrupted task [3, 37, 87, 154]. Clifford & Altmann [45] showed that asking people to manually make more complex cues had a negative effect on performance, however Iqbal & Horvitz [105] demonstrated that automatically generated cues could have a positive performance benefit. Reminders will be discussed in more detail in Section 2.4.2.

Another factor that could play a role that was not directly considered by Miyata & Norman [152] is the environment. Grundgeiger & Sanderson [75] reviewed the literature surrounding the effects of interruptions in a hospital environment. The aim was to discover the positive and negative effects that such interruptions would have; however, the authors found that the quality and scope of existing work was not sufficient to identify the effects of introducing notifications into a hospital (or similar) environment and called for further research to be carried out. This literature review was unusual because the authors limited their sources to several prominent American journals, ignoring a large number of seminal works on the effects of interruptions. However, the paper does highlight how little domain-specific work has been carried out into interruptions in certain environments. While a great deal of work has been carried out in cars, most studies into the effects of interruptions are based in the lab. A hospital is a dynamic multi-user environment with a great many users and various different devices which (presumably) move around the hospital. Despite the limited source material, Grundgeiger & Sanderson are correct to point out that existing work fails to adequately explore the relationship between the environment and the effects of the interruption.

Thesis Question 2 considers the effects on the user of delivering notifications using different methods. This section has shown that the effects of notification delivery are generally quite well understood, and that well-designed notifications delivered at the appropriate time can minimise negative effects while providing the desired benefits.

However, it is not clear if these effects are consistent when the notifications are delivered in different modalities, as suggested by Perry *et al.* [171], and more work is needed to fully answer this thesis question.

2.4.2 Designing Reminders

Interruptions in certain environments, such as an office or airplane cockpit, are usually designed with the surrounding environment in mind. However, in the field of human-computer interaction this environment is often an unknown factor. For example, Miyata & Norman [152, p. 227] argued that an ideal reminder should satisfy the following five conditions:

- inform the user when conditions are ready for resumption of a suspended or backgrounded activity;
- remind the user when something has to be done immediately;
- not distract from the current activity;
- continuous or periodically list activities that have been suspended or backgrounded;
- help resumption of an activity by retrieving the exact previous state of the activity and making it available to the user.

However, like much of the work in this field, Miyata & Norman are chiefly interested in multitasking in the workplace. In the home, continually listing activities that have not been completed would likely become irritating, especially if the reminders concern leisure activities which are not particularly important.

Ju & Leifer [112] point out that a human will interpret the context of the situation in order to be considerate, and that by not doing so, technology has become obnoxious. The authors use the analogy of a hotel doorman and compared it to an automatic door. They argued that behaviour that is acceptable from an automatic door would be considered extremely rude from a human doorman, such as not indicating whether the door is locked. This example was extended to show how an automated doorman could be improved by adding simple indicators of intent in lieu of human interaction, providing the additional 'polite' behaviour that is expected when humans interact. Bickmore *et al.* [22, 23] found a higher compliance for polite reminders than firm or quick ones, showing that 'polite' computers are more likely to be accepted by users. Arroyo *et al.* [10] found that polite reminders were much more effective in a driving simulator at reducing driver errors.

Gibbs [70] agrees that computer-based interruptions are inconsiderate; that screensavers will come on during presentations and mobile phones will ring in movie theatres. However, he claims that part of the problem is that computers are still generally unaware of their context. Gibbs discusses how interruptions can be managed by rules and models, allowing them to make a 'best guess' attempt to infer context from a limited number of information. A well-known example of this is a spam filter, which uses a system of rules to remove unwanted e-mails from on the behalf of the user. Such systems have proven very successful in the past, but are still regarded with caution; many users are unsure of their inner workings and worry about important messages or interruptions being suppressed.

McFarlane & Latorella [141] define four models for managing interruptions, and argue that coordinating interruptions can maximise their benefits and reduce their negative effects. Immediate interruptions are delivered whenever the system decides to do so (such as Internet pop-up advertisements) are decried as being responsible for most of the negative effects associated with interruptions [70, 112].

Negotiated interruptions give more power to the user, offering the choice of how and when to handle them (such as the 'ask me later' box for Windows updates). The FEELIM system developed by Espinoza *et al.* [61] is a good example of a system that allows the user to negotiate with the interruption systems.

Mediated interruptions use information about the user and their activities to determine how 'safe' it is to interrupt the user. For example, the BusyBody system developed by Horvitz *et al.* [96] attempted to avoid interrupting a person when whey were busy with another task. Monk [153] showed that the timing of interruptions plays a role in their disruptiveness; by moving interruptions out of an activity's early planning stage, the interruptions became much less disruptive. Other work by Horvitz *et al.* [93] shows that users cycle from busy to available quickly, and that a short waiting time during busy periods can reduce the overall disruption caused by interruptions.

The final way to reduce the negative impact of interruptions is through explicit scheduling; the user will know when an interruption will come, and can make the most of it. The problems with this model are immediately obvious; why would someone need a notification if they know when it will come, and what purpose would it serve? McFarlane & Latorella cite a paper by Hall & Hursch [76], where time management training was given to a University Physicist. The training allowed the physicist to organise when they would be interrupted by their students, allowing more time for high priority tasks.

McFarlane & Latorella point out that a fixed schedule, such as a school bell, allows for better allocation of time than an immediate interruption system. Of course, this model is only applicable in some areas; in many cases interruptions cannot be planned for in this way.

As discussed in Section 2.1.3, there are many complex requirements when developing technology for the home. While negotiated interruptions might give people control in their home, research suggests that mediated interruptions might be the most acceptable way to deliver reminders in the home [203, 202]. This could be used to make home care technology more appropriate, and should be considered when exploring research questions 3 and 4.

2.4.3 Scheduling Reminders

One of the most well-known theories for interruption deferral is the Attentive User Interface, first proposed by Horvitz *et al.* [94]. Horvitz *et al.* proposed that observations about the user's locational and activities could be used to infer the current activity of the user, and that some activities were more 'interruptible' than others. Horvitz *et al.* suggested that notifications could be deferred if the cost of deferral was lower than the cost of interrupting the current activity. This work was later expanded to make personalised deferral decisions based on machine learning principles [92], voice and facial recognition input [95], mobile phone input [97] and information volunteered by the user themselves[30] [96].

McCrickard & Chewar [140] argued against the Attentive User Interface, stating that it fails to consider the user's goals in relation to the notifications. While Horvitz prioritized notifications when working out the cost of deferral, McCrickard & Chewar argue that the cost must be worked out in relation to the user's current and overall goals. McCrickard & Chewar even went on to argue that different styles of presentation (in particular, different levels of saliency) should be used to minimise the intrusiveness of notifications that do not help the user achieve their current goals.

Another school of thought on interruption scheduling is breakpoint theory. This is based on a theory suggested by Miyata & Norman [152]; that the best time to interrupt a person is between tasks or subtasks. Based on this idea, Bailey & Konstan [11] demonstrated that the negative effects of interruption could be significantly reduced

[30] In this case, the users were asked when they were busy and could set timers that prevented interruption until they had expired.

by aligning the interruptions with natural task boundaries. Iqbal & Bailey [104] later expanded on this work with the creation of the OASIS framework. The OASIS framework is based around the idea that some breakpoints are easier to infer than others; *e.g.* the change from programming to media playing represents a *coarse* breakpoint, while switching from programming to debugging is a *fine* breakpoint [104, p. 15:6]. Iqbal & Bailey's work focussed primarily on the multi-tasking situations, but Bogunovich & Salvucci [26] were able to demonstrate that breakpoints could also be detected in single-task activities. This means that breakpoints could be predicted based on user behaviour, instead of simply watching for them to occur and attempting to capitalise on them.

Both Breakpoint theory and the Attentive User Interface theory have primarily been tested in the computer/office context, yet both could also be applied in the home. However, both theories fail to take into account the 'human aspect' prevalent in the literature discussed in Section 2.4.2. Jorge [111] argued that technology for the elderly should provide 3 basic functions for the best standard of living: provide non-intrusive observation, support daily routines and enhance social communication. Are there important differences between interrupting a social situation compared to interrupting a business situation? Bardram & Hansen [14] argued that this was the case and created the AwarePhone, which gathered the user's self-declared status (busy, free, etc.), schedule and location and made it available to anyone who wished to contact that user. This information would allow a person to use their understanding of social situations to decide whether to interrupt the user or to leave a message instead. However, given the range and depth of human social interaction it seems that a computer able to automatically detect and interpret complex social situations in the home is currently quite far away [57, 112].

Vastenburg *et al.* [203] carried out a field study of a home-based notification system. Their system was originally based around the Attentive User Interface, which attempted to ascertain a good time to interrupt the user. Several reminders were created with a home-style context and split into 3 levels of importance. A dynamic system would then decide whether to interrupt the user based on how important the message was and the user's activity. The results of the study showed that messages mediated in this way were considered highly acceptable. Vastenburg *et al.* [202] followed up this work by carrying out another study in a live-in laboratory. This time, Vastenburg *et al.* also changed the presentation of the messages; specifically, how salient they were. Vastenburg *et al.* found that managing the salience of the messages had a much larger effect on acceptability than controlling the timing, and that the message's importance was the primary measure of acceptability. The following section will explore different ways of presenting reminder notifications and their potential applications.

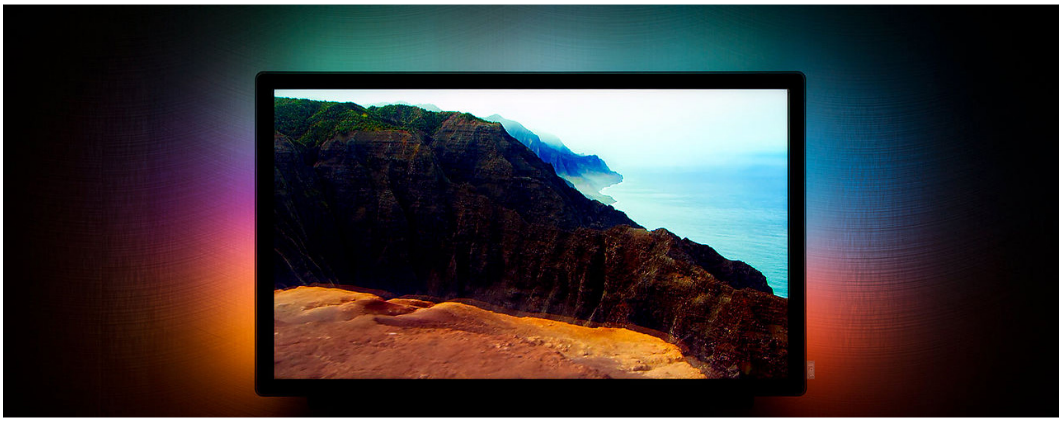
2.4.4 Reminder Modality

In the context of HCI, modality usually refers to a specific method of information transfer between the computer and the user (and *vice versa*) such as text, speech, *etc*. Technology which uses a single modality to interact is called unimodal, and technologies which use multiple modes is called multimodal. Caprani *et al.* [41] argued that the primary shortcoming of the Cook's Collage project (discussed in Section 2.3.2) was a failure to tackle sensory impairment and suggested that multimodal interaction would be idea to approach this problem. Multimodal technology has been advocated in this capacity several times by many other researchers [59, 108, 143, 162, 171, 226, 228]. This section will explore the different modalities that are used for interaction, the types of communication they support and what can be achieved when they are combined.

Visual Display

As a great deal of human communication is carried out visually, so there are several ways to display information visually. The most common forms of visual interaction are undoubtedly through writing and pictography. The most common forms of visual interaction in HCI are through text and pictograms.

When sending notifications, text allows a great deal of information to be presented. However, it can take time to convey its meaning and requires that the targets are literate in the language used. The target must also give their attention to the area where the text is displayed. Textual representations can also include numbers, which can help to provide background information. Adams & Edworthy [1] studied how the presentation of text can change the way information is transmitted and perceived, noting that perceived urgency was related to the text size, weight of the border and colour; white space was also tested but not found to affect urgency. Khan & Lenk [117] compared the typography standards for print and found that they were not the most effective for the display of text on a screen; new standards for presentation were needed. McCrickard *et al.* [139] tested various ways to animate text to improve salience and information transfer. Text can be presented in a great many different ways, and the guidelines change depending on the medium and the intent of the message. While the presentation of text is a research area too large to fully explore in this literature review,



(a) The Philips Ambilight.

(b) The Microsoft Illumiroom. Note that the projection is distorted to 'flatten' furniture from the perspective of the user.

Figure 2.8: Two examples of ambient visual interaction devices made by Philips and Microsoft designed for the home environment.

it is clear that text is an effective notification medium in many settings, *e.g.* warning label designs.

Pictograms are visual representation of objects, actions and ideas. They tend to be limited to transmitting only simple information, but they have several other advantages; they are simple, do not require literacy and can transfer their meaning quickly. Young [227] found that warning labels were more likely to be noticed and were understood quicker when accompanied by colour pictograms. There are several ISO documents entirely devoted to pictogram design, particularly signs [106] and warning labels [107]. Like text, pictographic interaction depends on the target focussing their visual attention in the right place. Pictograms are also harder to change than text, as a new pictogram must be designed if a suitable one is not available. Pictograms are also open to interpretation, and might be interpreted differently depending on the context. The study of pictograms is outside the scope of this thesis; it is enough to simply acknowledge their applications.

Ambient visual displays tend to refer to technology that delivers information to the user via the surrounding environment. This technology often uses peripheral vision and simple abstract messages to deliver information. An example of a commercial system is the Philips Ambilight [216], shown in Figure 2.8a, which projects colour on the wall around a television to improve immersion and set the mood. Microsoft's IllumiRoom [110] is another example, which attempts to extend the field of view of a television by projecting images onto the wall behind it as shown in Figure 2.8b. Arroyo & Selker [7] reported that simple light-based ambient displays were quite effective as notifications. Weiser & Brown [218] called for 'calm technology', technology which "empower[s] our periphery" [218, p. 2]; one of the examples given is a piece of string which twitches to show load on the computer network. It is likely that ambient visual technology will be highly appropriate for the home, as it is designed with the environment in mind.

Audio Interaction

There are several ways to interact using sound. When sending messages, McGookin & Brewster [148] claimed that there are four primary methods; sonification, speech, auditory icons and earcons.

Sonification is a way to use audio to describe trends in data. There are several natural instances of this, *e.g.* gauging the speed of a car by the sound it makes when it drives past. Sonification is also used in technology to simplify the process of understanding data; a good example is a Geiger counter, where faster clicks signifies higher levels of radiation without giving absolute values. Sonification has been used to present several different types of information including graphs [35, 205], tables [35], spreadsheets [191] and meteorological data [172]. Sonification might be be useful when trying to present information to visually impaired older users, *e.g.* sonified graphs of activity.

Speech is an effective interaction medium as it is one of the natural ways in which people communicate. Unlike other auditory methods, speech is able to deliver absolute numerical data. Using pre-recorded human speech is resource intensive, as either sentences, words or syllables must be recorded; if sentences are used, there is little space for interactivity, and if smaller speech 'chunks' are used they often sound unnatural when reassembled into sentences. Synthetic speech generated by a computer is an alternative which has been criticised for sounding 'robotic' and 'unnatural' in the past, although modern systems able to generate much more natural sounding speech that is even able to emulate an accent [146]. The interpretation of speech can also be controlled; Hellier *et al.* [86] found that the perceived urgency of a speech warning message changed with the phrasing and tone of voice. Several researchers have argued that speech should be one of the primary ways to interact with older people as conversation is both natural and familiar [73, 145, 168, 224, 229].

Auditory icons are the aural equivalent of pictograms; short sounds which hint at their meaning. Another definition is "everyday sounds mapped to computer events by analogy with everyday sound-producing events" [68, p. 1024]. Gaver [67] noted that that main problem with auditory icons lies in creating them, as they must be recorded from real life. Gaver proposed a way to synthesize some auditory icons by modelling the interactions between materials but was limited to simple sounds such as bumps, scrapes and impacts. Auditory icons are intended to be quicker than speech to interpret while remaining easy to understand, which could make them an effective notification medium. However, as they do not fully obscure their meaning interception of messages is a possibility. Auditory icons also rely on the receiver understanding the relationship between the message and the sound, making them somewhat subjective; it is also true that for many messages, a suitable sound to describe it will simply not exist.

Earcons are abstract, structured synthetic sounds that work much in the same way as auditory icons. The main differences are that earcons, being abstract, do not hint at their meaning; the associations between the sounds and their messages must be learned. Another important difference is that earcons can be fully synthesized by a computer without any need for recording equipment [148]. Brewster *et al.* [32] created guidelines for the creation of earcons; the primary parameters were timbre, register, pitch, rhythm and intensity. Edworthy *et al.* [58] and Brewster *et al.* [32] identified ways to change the perceived urgency of an earcon. However, Arrabito *et al.* [5], noted that interpreting earcons can be difficult; in their case study of helicopter warning sounds, subjects were generally able to identify the urgency of the earcons but there were issues identifying what triggered the alert. As the interpretation of earcons must be trained, this might make them less suitable for use in the home when compared to auditory icons or speech.

McGee-Lennon *et al.* [147] proposed the idea of Musicons; very short snippets of songs that serve as notifications. It is suggested that Musicons provide the same benefits as Earcons, being private and abstract, yet would be more memorable; in particular, it is suggested that associations could be made between the lyrics of a song and the message being delivered. Musicons were compared to speech notifications and were found be highly effective. However, this work is quite new and McGee-Lennon *et al.* note that more research is needed to fully understand how musicons can be created and deployed effectively.

Tactile Interaction

Tactile interaction is interesting because the human sensation of 'touch' is actually the product of several sensory systems with their output combined by the brain. While a full overview of the somatosensory system is not required, the primary system that make up a sense of touch are as follows:

- Nociception, a system comprised of a network of nociceptors throughout the body that sense damage and are responsible for the sensation of pain;
- Proprioception, a system comprised of receptors in muscles and in the inner-ear which create the sense of self-position and balance;
- Thermoception, comprised of two different sensors which detect heat and cold, giving the ability to detect relative temperature;
- Mechanoreception, a heterogeneous network of receptors that is responsible for detecting pressure and vibration.

In HCI research, tactile interaction generally considers interaction with the mechanoreception system via vibrations. Information delivered in this way is typically called a tacton. Brewster & Brown [30] noted that when designing tactons, some of the parameters used were the same as those used when designing earcons. Different tactons could be created by changing the frequency, timbre, amplitude, waveform, duration, rhythm and body location. Hoggan & Brewster [90] compared tactons and earcons that both varied their roughness (by changing the amplitude) and rhythm by the same amounts; *i.e.* the earcons and tactons were directly comparable. The results showed that tactons were effective at delivering information. Tactile devices are readily available given their ubiquity in modern mobile phones, which already use simple tactons to inform the owner that a text message or a phone call has arrived. This technology is usually based around a vibration motor, which is a motor with an asymmetric weight attached. Such devices have a slight delay and a short 'spin down' after the message ends, and do not have the fidelity to deliver complex messages. Devices such as the C2 actuator[31] are an off-the-shelf solution able to deliver much more complex and precise tactile messages.

Temperature based interaction is rare, but not unusual. Arroyo & Selker [9] used a mouse augmented with a peltier to interrupt a task, noting that it was as effective as other methods tested; although in a later experiment they claimed it was inferior to ambient visual interruptions [7]. Wilson *et al.* [223] considered the possibility of creating 'thermal icons' by rapidly changing temperature. Wilson *et al.* noted that the identification rate for thermal icons was high, possibly even higher than that of tactons, suggesting that thermal interaction may be effective. Peltiers are also completely silent and rely on no moving parts. However, the technology used by Wilson *et al.* [223] and Arroyo & Selker [9, 7] was custom-built; consumer-ready thermal interaction devices do not currently exist.

Pain is an interesting concept for HCI; Eccleston & Crombez [56] noted that pain will override all other stimulus when interrupting due to the 'threat factor'. Yet what is a pain-based-interaction device, and who would agree to use it? Vaucelle *et al.* [204] designed a haptic device that caused pain: an inflatable bladder with spikes that pressed into the arm. The device was intended to be used as a therapy aid for people who self-harm, and a clinical trial of the device demonstrated that it was effective in its role. Although pain may be a powerful motivator, it is clear that pain-based interaction raises serious ethical concerns and it is unsuitable for consideration in a home care scenario.

Alternative Interaction Modalities

Olfaction and gustation are two interaction opportunities that are generally overlooked in HCI research, most likely due to issues surrounding their practicality. Olfaction is particularly interesting as it is has ties to several areas of memory, suggesting it could be useful as a reminder. Brewster *et al.* [31] investigated this in an experiment that attempted to link smells to photographs. While the experiment found that some participants could use scents to recall photographs, it was not found to be particularly effective when compared to text. Brewster *et al.* also reported several practical issues when trying to store, deliver and dissipate smells. Bodnar *et al.* [25] considered the potential of smell as an ambient notification system, finding that while smell was less effective, it was also less disruptive. Bruck & Brennan [36] investigated which cues of a

[31] Engineering Acoustics, `http://www.eaiinfo.com/`.

fire at home would be most likely to wake a sleeping person, which included smell. A significant gender divide was discovered in smell, with 80% of females waking on the smell of smoke compared to 29% of males. There are several obvious issues with smell; it relies on perishable chemical components, it is difficult to control and highly subjective. However, relatively little work has been carried out into its potential applications as a notification medium.

Interaction through taste is extremely rare, although there is one notable example in BrainPort [52], a brain-computer interface device that communicates by stimulating the tongue. The device is intended to help blind people to navigate their environment, and tests have shown that the technology can be quite effective. However, in the context of home care it is unlikely that a gustation-based interaction device would be acceptable or practical.

Comparing Modalities

Thesis Question 1 asks which of these modalities are appropriate for use in a care scenario. There are several ways that a home care system could interact with a user in the home, each with their own potential applications areas. This section reviews literature that compared modalities to one another in an attempt to clarify which modalities would be the most effective and appropriate for a home care setting.

Latorella [123] performed a study that compared the disruptiveness of visual and audio reminders to visual and audio tasks in airplane cockpits. Latorella made several interesting observations, in particular noting that audio interruptions seemed to interact with an audio task to create a much higher number of errors, while the other combinations were were not significantly different. Latorella also noted that participants appeared to give priority to audio tasks over visual interruptions, which was also observed by Nikolic & Sarter [164] in a similar experiment. Latorella's findings show that the response speed and disruptiveness of an interruption is influenced by the modality of both the original task and the interruption.

The most broad comparison of the modalities was carried out by Arroyo *et al.* [9], who investigated five modalities: heat, smell, sound, vibration and light. Participants were asked to carry out some tasks that included counting backwards and reading text, and would be interrupted by one of the modalities which was a signal to stop the primary task and carry out some secondary task. Disruption was measured in terms of reduced comprehension, mistakes made and increase in completion time. The results of the experiment were not statistically significant and they were unable to show that any one modality outperformed the others. Arroyo *et al.* also administered a subjective survey which demonstrated the existence of a *perceived* level of disruption. While this survey does appear to show that subjects perceived a difference in their performance, the survey was flawed in that subjects were asked to rank-order the modalities instead of rate them on a scale. Based on this data Arroyo *et al.* concluded that the users were most disrupted by the interruption methods they were unfamiliar with.

Arroyo & Selker [8] went on to run another experiment comparing heat and light, attempting to demonstrate that performance and perceived disruptiveness differs with the interruption modality. Subjects were asked to play a complex text-based adventure game, and disruptiveness was measured as the number of mistakes made post-interruption. Their results demonstrate that interrupted games had significantly more errors than uninterrupted games, but were not able to fully demonstrate a difference between different modalities. The authors were also unable to correlate preference and performance, as the subjects' expressed modality appeared not to be linked to the modality which offered them the greatest performance in the game.

McGee-Lennon *et al.* [146] compared the properties of 3 different types of audio notification: speech, earcons and simple beeps. The aim was to find out which of the 3 modalities were the most effective, disruptive and preferred. The authors concluded that speech seemed to be most effective and least disruptive, yet also found that many participants preferred earcons. McGee-Lennon *et al.* noted that there did not seem to be a relationship between the modalities that a user liked and the modalities that were most effective, which was also reported by Arroyo & Selker [8] in their conclusions.

Emery *et al.* [59] carried out an experiment into multimodal feedback for a drag-and-drop task with older users who had different levels of technology experience. Visual, audio and tactile feedback were tested individually and in combinations to find out which were most effective at reducing the error rate in the task. The results showed that combinations of modalities would outperform single modalities, with three modalities generally providing little additional benefit compared to two modalities. Emery *et al.* also noted that visual feedback seemed to provide the poorest performance, and suggested that this was due to the way older adults process visual information. A related experiment by Jacko *et al.* [108] tested the same modalities with older users with and without visual impairments; the results showed that visual feedback was poor for all groups, not just the impaired group. Jacko *et al.* also found that the impaired users received a much larger benefit from non-visual feedback than the unimpaired users.

Hoggan *et al.* [91] examined audio and tactile feedback to investigate when they should be deployed, and found that a high volume of background interference in the same modality would lower the effectiveness of the feedback. Specifically, background noise would lower the effectiveness of audio feedback and vibration would lower the effectiveness of tactile feedback. Hoggan *et al.* suggested that devices could attempt to detect this background noise and automatically switch to a less busy sensory channel.

Lee *et al.* [128] compared tactile and auditory feedback in a collision warning alarm for use in a car. Their results found no significant differences between the audio and tactile alerts, but did find that grading the feedback based on the urgency of the alert had a significant impact on the car's deceleration.

Bodnar *et al.* [25] compared olfactory notifications, earcons and abstract visual messages with the aim of evaluating the smell-based interruptions. Their results showed that olfactory notifications reduced task completion, but did not result in a lower error rate in the task. The olfactory notifications were also less effective. However, their subjective survey seemed to show that participants found the olfactory notifications to be less disruptive than the other modalities.

McGee-Lennon *et al.* [145] carried out a large survey ($n = 379$) of older users that investigated why they wanted reminders, their current reminder solutions and what they expected from reminder technology. This resulted in some interesting observations. It was found that people do not discount impaired senses when considering which modalities they would like to interact in (it is assumed that this does not apply in the case of totally impaired senses). Their results also suggested that people wanted to interact in different modalities depending on the device in question and their age. Smell proved particularly unpopular, with only 3% of people preferring it to other methods. The results of the survey clearly suggest that the acceptability of the modality chosen to deliver a reminder depends on several factors including personal preference, current activity, sensory impairment and social situation.

These studies tend to show that while there are differences between modalities, there is no clear 'best' modality to use when delivering a notification. There are several important factors to take into account, such as the modality of the task being interrupted [123], the number of modalities used to notify [59, 108], the preferences of the target [8, 146, 145], interference in that sensory channel [91] and sensory impairment [146, 108].

There are also some neurological factors that might influence modality choice that are not well understood. For example, while certain parts of the brain might appear to have a specific purpose (*e.g.* processing audio stimuli) Laurienti *et al.* [124] have shown that some resources can be 'reallocated' during ongoing activities. Laurienti *et al.* specifically refer to parts of the brain that would normally process audio being active during an ongoing visual task, providing a boost to visual processing but reducing the ability to respond to audio stimuli (and *vice versa*). This property is known as cross-modal plasticity [16], where the brain will reconfigure itself over time to strengthen non-impaired senses. Such neuroplasticity is likely to explain why Jacko *et al.*'s [108] impaired participants received a greater benefit from multimodal feedback in their non-impaired senses compared to unimpaired participants. Laurienti *et al.*'s findings suggest that a similar process appears to happen very quickly in prolonged visual and audio tasks, which suggests that time spent engaged with a unimodal task will affect the ability to receive reminders in different modalities. As human perception is based on the synthesis of data from various sensory channels, data from one sensory channel can affect the perception of data from another channel [185]. A well-known example of this is the McGurk effect [149]. While it is important to understand the existence of this effect and to consider the implications for multimodal interaction (particularly for experimental design), understanding the neurology of sensory perception is outside the scope of this work.

Considering Thesis Question 1, it is clear that there are many factors which play a role in determining which modality (or modalities) should be used to deliver home care notifications. Many of these factors are emergent properties that cannot be predicted, and as discussed in Section 2.1, many of the factors change over time. This has led to a common theme in research which attempts to compare modalities: that multimodal technology should be smart enough to select the 'best' modality in real time, instead of relying on one or two modalities to fulfil every need [8, 9, 91, 145]. Thesis Question 3 aims to form a better understanding of this.

Dynamic Notifications

Section 2.4.3 discussed several ways to manage the timing of interruptions to minimise their negative effects. Section 2.4.4 has shown that there are also several ways to manage *how* a notification is delivered, and many researchers have called for systems that are able to switch between modalities and configurations to suit the situation.

There are some interesting examples of smart dynamic notification technology in research. Arroyo *et al.*'s [10] CarCOACH technology combined multiple alerts to reduce stress and used 'polite' interactions to great benefit. Lee *et al.* [128] showed that varying the urgency of a collision alarm resulted in better braking. However, there are few examples of this sort of technology in the home setting.

McBryan & Gray [137] discussed some of the ways that home reminder technology could switch between interaction modalities using various techniques. One method suggested was that the user's preferences could determine which device (selecting from phone, television and a loudspeaker) is used. However, McBryan & Gray believed that this approach would not be appropriate in every circumstance, and that users might disable some output options to manage the system's behaviour. McBryan & Gray argued that context sensitivity would play an important part in deciding which device to use for a notification.

Vastenburg *et al.* [202] tested a system that changed the salience of a notification based on how important the message was and found that this resulted in both effective and acceptable interruptions in a living-room lab experiment. Perry *et al.*'s work [171] on the millennium home (see Section 2.3.1) is also important, advocating for mode-switching technology and outlining several important design issues. In general however, technology that actively controls the modality or configuration of notifications is rare, and it is not well understood what the implications of mode-switching in the home would be, or the best way to design such a system.

2.4.5 Observations on Reminder Technology

The work presented here demonstrates that there are many modalities which can be used for interaction, and inside most of those modalities, a wealth of configuration options to change both the message itself and to emphasise certain aspects of the message (*e.g.* several methods have been discussed to make a message appear more urgent). This means that answering Thesis Question 1 is not straightforward, as there are several modalities that would be suitable for use in a home care system depending on the context. It's also clear that further research would be needed to answer Thesis Question 2, as while the effects of interruptions on a person are quite well understood, the role that modality plays (if any) on those effects is not.

Thesis Question 3 considers how a system can be designed to best utilise a range of notification delivery methods. While many researchers have advocated for intelligent and dynamic multimodal technology able to manage interactions and promote effectiveness and acceptability, there are relatively few examples of this sort of technology in research. Creating such technology would rely on a good understanding of how to design notifications in each of the desired modalities. A good understanding of the relative strengths and weaknesses of each modality would also be needed. While a lot of good work currently exists on these subjects, in order to fully understand when and why to apply a given modality, a more comprehensive overview is needed that is based around the home reminder context.

2.5 Summary & Conclusion

Thesis Question 1 asks "which forms of interaction are appropriate for use in a home care system?". Section 2.1 showed that home care technology users are likely to suffer from sensory impairment [201] and cognitive decline [169, 179]. Multiple forms of interaction can be used to address sensory impairment by using non-impaired channels. Section 2.2 explored the home care industry and showed that interaction with the user is not the norm, although some companies are moving towards involving the user with multimodal interaction devices [200]. Section 2.3 examined research projects, which revealed that many researchers were advocating for user inclusion through notifications designed to improve the general standard of living [17, 22, 54, 100, 125, 157, 163, 171]. Section 2.4 examined different types of interaction and identified several modalities that could be used to interact. Many of these modalities are uncommon yet could be used to solve some of the unique challenges found in home care situations, *e.g.* olfaction can be used to communicate with a deaf/blind person without requiring physical contact. Existing research that compares these modalities does not conclusively show which modalities can be used to deliver home care notifications, and as such further work is needed to form a better understanding of how these modalities can be used in home care scenarios.

Thesis Question 2 asks "how do different forms of notification delivery affect users?". Section 2.4 showed that there is a wide range of potential negative effects that can be produced when interrupting a person engaged in a task [12, 39, 42, 46, 134, 180]. However, there is relatively little research that looks into the role that modality plays, in particular whether some modalities are more disruptive of distracting than others. There's a chance that more distractions would pose a risk to the user, *e.g.* Berg *et al.* [21] showed that distractions were one of the primary causes of falls in elderly people at home. Further work would be needed to fully understand how effective and disruptive notifications are when delivered in different modalities. That information would be key to answering Thesis Questions 3 and 4 on how to create multimodal reminder technology is appropriate and acceptable for the home context.

Thesis Question 3 asks "how can home reminder technology be designed to best utilise multiple types of interaction?". Several researchers have advocated for home technology simple enough to be installed and configured by the end user [55, 57, 85, 99], although Sections 2.2 and 2.3 have shown that the modular designs common in research and industry often do not meet this goal. Modular designs are common due the the need for flexibility in home care, as the home environment and user's needs are likely to change over time [57, 62, 137, 207]. Another way to improve the flexibility of home reminder technology is to make it more intelligent using techniques like interruption management. Interruption management systems generally aim to reduce disruption from interruptions by controlling *when* a notification is delivered to users [94, 104, 152]. Several researchers have argued for technology that is also able to control *how* notifications are delivered [9, 8, 137, 171], and research has shown that this can be an effective way to improve the appropriateness of the technology [10, 203, 202]. Additional research work is needed to find a design for home care technology that is able to use multiple modalities, in particular technology that is able to control *how* and *when* to interact to minimise disruption and maximise the appropriateness of notifications.

Thesis Question 4 asks "can home reminder technology be made more effective and appropriate by providing it with the ability to dynamically select from multiple forms of interaction?". Section 2.2 has shown that one of the major shortcomings of industry products are that they provide no salient advantages to the care recipient, generally focussing on providing piece of mind or observational tools for carers and family members. While this technology might allow a person to remain at home who would otherwise require residential care, Beaudin *et al.* [17] found that it was likely to be seen as intrusive unless it interacted directly with the care recipient. Section 2.4 noted the trend towards reminder technology, and there is evidence that reminders can provide more appropriate and effective care than technology that does not interact with the user [100, 125, 163]. Given the prevalence of sensory impairments in home care technology, it seems logical that multiple forms of interaction would make the technology more appropriate and effective. More research would be needed to fully understand if multiple modalities would make home care technology more appropriate and effective for those without sensory impairments. Work is also needed to fully understand how switching between modalities would affect the interactions between the system and the user [171].

This literature review has shown that there are several ways for home care technology to be improved. This thesis explores how multiple forms of interaction can be coupled with dynamic technology to provide more appropriate and effective home care technology. Chapter 3 addresses Thesis Question 1 by evaluating several forms of notification delivery to determine which methods are appropriate for use in home care technology. Chapter 4 expands on the findings of Chapter 3 to address Thesis Question 2, exploring how interruptions and distractions in different modalities would affect an ongoing task. Thesis Question 3 is explored in Chapter 5, which presents the design of a dynamic multimodal reminder system called Dyna-Cue that is able to control *when* and *how* notifications are delivered. Chapter 6 used the Dyna-Cue prototype to address Thesis Question 4, clarifying whether this type of dynamic multimodal reminder technology is able to provide a more appropriate and effective solution compared to current forms of reminder delivery.

Chapter 3

Baseline Study of Notification Delivery Methods

Thesis Question 1 was *"which forms of interaction are appropriate for use in a home care system?"*. The literature review in Chapter 2 identified several modalities that can be used to deliver notifications, but noted that more research is needed to fully understand which ones should be used for the provision of care. This chapter presents a study carried out to gather baseline data on the performance of notifications delivered to different sensory channels.

This chapter makes a distinction between *sensory channels* and *modalities*. A sensory channel is a specific pathway for information, *e.g.* the visual sensory channel. Modalities are specific ways of delivering information, *e.g.* text or pictograms. While the later chapters of the thesis will consider individual modalities, this chapter presents an evaluation of the underlying sensory channels.

The aims of this study are laid out in Section 3.1, followed by Section 3.2 which identifies the design of the study. Section 3.3 extrapolates the aims into specific, testable hypotheses. Section 3.5 describes the procedure used, followed by the results which are presented in Section 3.6. The results are discussed in Section 3.7, refined into guidelines in Section 3.8, and finally the conclusion is presented in Section 3.9.

3.1 Aims

The overall aim of this study was to provide exploratory baseline data on the potential of different modalities to function as notification delivery mechanisms. This involved evaluating a range of modalities in a way that would allow a fair comparison of their properties despite their differences.

As discussed in Section 2.4.1, interruptions have a disruptive effect on ongoing activities. Some studies have suggested that the sensory channel used [8, 123] or modality [146] of an interruption will affect the amount of disruption to the interrupted task. However, other studies have found no clear difference in the disruptiveness of different sensory channels [9, 25, 128]. As distractions have been linked to falls in older people at home [21], it is important to have a good understanding of the disruptiveness of notifications delivered to different sensory channels. Evaluating the disruptiveness of notifications delivered to different sensory channels was one of the primary aims of the study.

Section 2.4.4 highlighted the wide range of sensory channels that can receive notifications, each with its own unique properties and mechanisms. The most thorough comparison of these was completed by Arroyo *et al.* [9], who tested temperature, smell, sound, vibration and light. However, Arroyo *et al.* did not present their data and their results were inconclusive. A more thorough analysis of the abilities of each sensory channel to receive information is needed, including performance data such as the speed of delivery. Research has also shown that speech notifications are responded to more quickly than visual notifications [123, 164]; more work is needed to extend these findings into a range of visual and audio modalities as well as to include alternative sensory channels in the comparison, such as tactile and olfactory. Evaluating the effectiveness of different sensory channels at receiving notifications is another aim of this study.

As shown in Chapter 2, one of the most significant shortcomings of current home care technology and research is a failure to consider the user. Subjective assessments should be carried out when evaluating the sensory channels to provide an insight into their suitability for use in the home. In addition, some research has shown that people may not prefer the modalities they perform best in [146]; therefore subjective performance assessments should be compared to actual performance measurements.

The final aim of the study was to explore the potential influence of Modal Learning Preference (MLP) on a participant's ability to process data received by different sensory channels. Modal learning theory holds that people will have a natural 'preferred' way of learning, *e.g.* some people might learn better by reading a book while others might learn better by listening to a lecture [65, 170, 184]. The preferred way of learning is called the Modal Learning Preference. There is no consensus on the taxonomy of modes that can be used for learning; the most popular models are VAK (Visual, Audio and Kinaesthetic) and VARK (Visual, Audio, Reading/Writing and Kinaesthetic). It may be that modal learning preferences will map to an improved ability to receive data in preferred modalities. The study also explored these issues.

The aims of the study can be refined into four research questions, as follows:

Research Question 1
: Does the sensory channel used to receive a notification affect how disruptive the notification is to an ongoing task?

Research Question 2
: Are some sensory channels more effective than others at receiving information?

Research Question 3
: What are the subjective assessments of the different sensory channels, and does perceived performance correlate to actual performance?

Research Question 4
: Does Modal Learning Preference (MLP) correlate to an improved ability to receive information in that sensory channel?

The following section describes the design of the study that was carried out to address these questions.

3.2 Design

The design of the study was based on experiments carried out by McGee-Lennon *et al.* [146] and Arroyo *et al.* [9]. Both experiments involved two activities: an ongoing activity which we call the *primary task*, and a short time-sensitive action that interrupts the primary task called the *secondary task*. For this study, the primary task chosen was a card-matching game which is described in the following section. The secondary task was to press one of three buttons, described in Section 3.2.2. Participants would receive notifications telling them to press a button and it was the responsibility of the participant to select the correct button. The modalities used in the experiment, along with their configurations, are described in Section 3.2.3.

The experiment itself was a repeated-measures design where each condition involved playing several card-matching games and receiving several notifications. The independent variables are specified in Section 3.2.4 and the dependent variables in Section 3.2.5. A discussion of potential confounding variables is given in Section 3.2.6.

3.2.1 Primary Task Design

There were two requirements for the primary task: it must represent a type of activity that would normally be carried out at home, and it must offer the ability to observe various behavioural changes. McGee-Lennon *et al.* [146] used a digit span test to evaluate serial recall in their experiment. While effective and well-validated, there is a limit in how much behavioural change can be extracted from this task. In addition, serial recall is unlikely to represent an everyday home task; this test may be more suited for an office context. The proof-reading test carried out by Arroyo *et al.* [9] was discounted for similar reasons.

Arroyo & Selker [8] used a complex computer role-playing game, switching from a proof-reading task to a leisure task. A leisure task (*e.g.* a computer game) might be more suitable for this experiment. In Arroyo & Selker's game users were expected to use their short-term memory and spatial reasoning to navigate and explore. Many home tasks (*e.g.* cooking and cleaning) also depend on short-term memory and spatial reasoning. While a leisure-based task would be ideal, this exact task would not be suitable for the experiment; Arroyo & Selker had problems interpreting their results due to the complexity of their task.

The activity selected for the primary task was a simple card-matching game based on the game 'Concentration', also known as Memory, Pelmanism[1], Shinkei-suijaku[2], Pexeso or Pairs. The rules are simple; pairs of cards are presented face-down to the player, who can then turn over two cards at a time in an attempt to find the pairs and remove them from the game. This game is useful for this task because it has been used to carry out psychological experiments on the memory development of children [13, 69, 181]. This

[1] Pelmanism was a correspondence course sold by the Pelman Institute in the early 20th Century. It was developed by W. J. Ennever and advertised as a way to improve memory. Source: `http://www.ennever.com/histories/history386p.php`

[2] Japan only

work shows how to use the game in an experimental setting and provides an insight into the measurements that can be taken (which will be discussed in Section 3.2.5).

Gellatly *et al.* [69] studied Concentration and its potential application in memory research. As part of this, Gellatly *et al.* identified two main strategies for playing the game: *primacy* and *recency*. The primacy strategy means when a player sees a card, and believes they have seen its match earlier in the game, the player tries to locate the earlier card first. A recency strategy means the user will click on the most recently viewed card first, then try to locate the card found earlier in the game. The primacy strategy will usually result in greater efficiency, *i.e.* fewer moves required to match all cards in the game. Gellatly *et al.* found that most players use the recency strategy regardless of age, despite it offering the poorest efficiency. As the game is linked to memory, and children tend to perform better at visual-spatial memory tasks than adults, it was believed that younger players would outperform adults. However, they found instead that performance does not appear to vary with age in players older than 9 years. This finding has been confirmed by other researchers, some of whom found that adults outperformed children [13, 181].

Schumann-Hengsteler [181] suggested that Concentration may not be entirely visual-spatial and theorised that adults have the ability to re-encode information on the cards. For example, a picture of a boat could be remembered by the picture itself, the verbal label 'boat', or simply by recognition of the object in the picture without explicitly labelling it. Schumann-Hengsteler suggests that this allows an adult is able to spread the mental workload over a range of resources, allocating the most powerful parts of memory to the task. This process furnishes adult players with a greater performance boost than the superior visual-spatial memory of a child player. However, this process cannot take place if the cards have abstract visualisations on them.

Concentration is a good primary task for several other reasons. It is a simple leisure activity that might well be carried out at home, it is a well-known game with very simple rules and it can quickly build a mental workload.

Configuration of the Primary Task

A game of concentration can be configured in many ways, including the type of card, number of cards, time allowance and layout. Schumann-Hengsteler [181] claimed that the content of the cards would significantly alter an adult's ability to play the game, so

(a) A French Deck. **(b)** Japanese Hanafuda. **(c)** Memory by MB Games.

Figure 3.1: Examples of the types of cards traditionally used to play Concentration.

care must be taken when selecting cards. The number of cards and the time allowed to find all the pairs was also important.

A traditional game of concentration would be played with a standard deck of 52 playing cards (known as a French Deck). Shinkei-suijaku, the Japanese version of the game, would be played with 48 Hanafuda ("Flower Cards"). Memory, made by MB Games, uses 72 picture cards designed to appeal to children. These are shown in Figure 3.1.

As well as these common variations on the original card game, there was also an American TV game show called Concentration. In its original run (1958-1973) the game was played with 30 cards, although in later shows this was reduced to 25 cards with a 15-card bonus round. The cards displayed pictures of the prizes which could be won by matching both cards. Odd-numbered games used a 'wildcard' which could match any other card.

As Schumann-Hengsteler [181] suggests that the player can re-encode the contents of the cards, the obvious choice would be to use pictograms. This would allow the player to re-encode the card's data as to the best of their ability. There are various ISO standard pictogram sets [107, 106], but they are generally used for sign-making and often lack easily identified verbal labels.

The game was instead configured using simple A-Z animal charactures taken from an online speech therapy resource website.[3] As they were designed for children, the pictures represent well-known items which would make it easy to translate into a verbal label. The entire collection of pictures is shown in Figure 3.2. Two images were omitted from the original set, which were the 'yo-yo' and the 'vest' pictures as they were slightly harder to interpret than the others.

[3]Speech Teach UK, http://www.speechteach.co.uk.

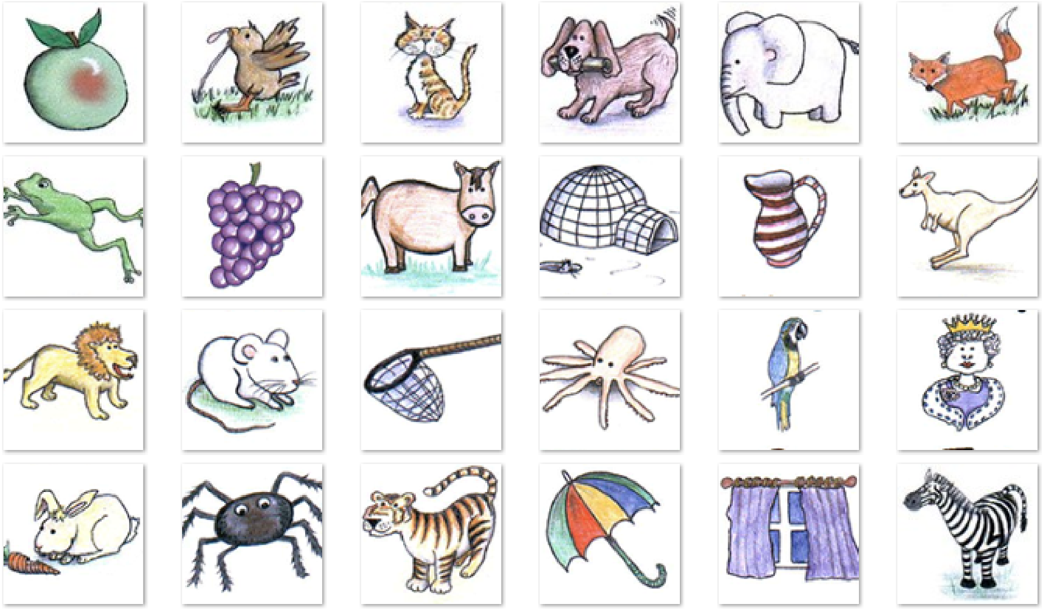


Figure 3.2: The animal-alphabet drawings used as cards in the Concentration game.

The experiment needed to be performed quickly enough that multiple trials and conditions could be completed, but there also had to be enough cards to ensure that the task was challenging. As shown by the wide range of variations on this game's premise, there is no standard configuration or number of cards. It is a common tactic to reduce the number of cards when a quicker game is desired; the Concentration television show reduced the number of cards to increase the speed of the game. The standard for research involving the game appears to be 32 cards [13, 69, 181].

The ideal configuration was determined to be a game that would take around one minute to complete, as it allowed all of the experimental tasks to be carried out in under 1 hour. To ensure that the game was correctly timed a rough prototype of the experiment software was created in Java. Some informal piloting suggested that 16 cards did not produce a challenging enough game but 24 cards took too long to play.

To answer questions surrounding the best configuration for the game the rough prototype was further developed into a Java Web Applet and advertised on the MultiMemoHome webpage. Users were asked to play three games of Concentration using 16, 24 and 36 cards. Participant information was logged to a database to allow simple data processing. The results, shown in Table 3.1, revealed that the original estimate was incorrect: the average completion time for 16 cards was actually less than half a minute. A 24 card game took an average of 63.9 seconds, very close to the minute-long game that was

Table 3.1: Results of a pilot study into the properties of different configurations of the game Concentration.

Cards	Arrangement	N	$\bar{t}$	$\overline{turns}$
16	4*4	43	41.36 ($\sigma = 2.9$)	15.9 ($\sigma = 0.5$)
24	4*6	38	63.89 ($\sigma = 2.8$)	27.5 ($\sigma = 1.2$)
36	6*6	37	125.01 ($\sigma = 4.1$)	53.6 ($\sigma = 1.8$)

Note: $\bar{t}$ was the average time taken to complete the game in seconds. $\overline{turns}$ was the average number of turns (each turn required two actions) taken to complete the game.

Figure 3.3: The three coloured buttons used in the experiment were labelled heating, lights and telephone.

desired. Using a 24 card game would mean that most players would either complete the game in just under a minute or run out of time close to the end of the game.

The pilot provided enough data to verify that the best configuration for the experiment was to use 24 cards arranged in a 6x4 grid. The cards were decorated with simple alphabet pictures as shown in Figure 3.2 and each game had a 60 second time limit.

3.2.2 Secondary Task Design

Gillie & Broadbent[71] found that the complexity of the secondary task and its similarity to the primary task were the primary factors influencing the disruptiveness of an interruption. As the aim of the experiment was to evaluate the disruptiveness of the modalities, it was desirable to minimise any effects from these two factors. That is, to isolate the disruptiveness of the different modalities, any disruptive effects from the secondary task itself must be kept to a minimum.

This was addressed by choosing a task that was extremely simple: pressing a button. A single button would not provide a way to measure notification comprehension, and is unlikely to reflect a home-based activity. Therefore, three buttons were used in

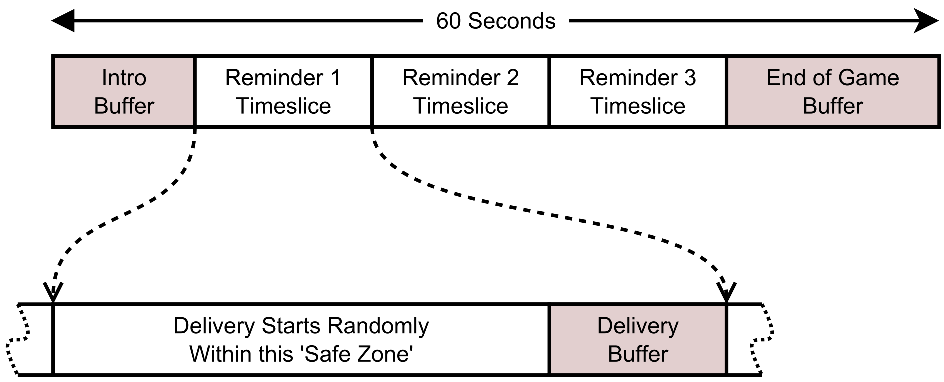


Figure 3.4: Diagram showing the division of time in a single experimental condition, including buffers at the start and end. Within a 'delivery timeslice', there is also a buffer at the end to ensure that notifications can finish delivery before the end of the allocated time period.

the experiment and were given a home-based context similar to McGee-Lennon *et al.* [146]. McGee-Lennon associated different sounds with a hypothetical home remote control, instructing the participant to turn up or down the heating, lights or a fan. These contexts are appropriate for this experiment, however it was difficult to create pictograms and auditory icons for 'fan'. Thus, 'fan' was changed to 'telephone' in the experiment, resulting in three buttons labelled 'Heating', 'Lights' and 'Telephone'.

As the participants would already be using the mouse, we considered the option of having them click a virtual button on the screen near to the game. Research has shown that physical boundaries have a distinct effect on task-related memory [173], so physically separate buttons were chosen to emphasise the separation between the primary and secondary tasks. This would also prevent participants from focussing too much on the on-screen display, which could result in the 'reallocation' phenomenon described by Laurienti *et al.* [124]. The buttons used were connected to the computer via USB and are shown in Figure 3.3.

For each of the given modalities there were three different messages, each corresponding to one of the three buttons. Where possible those messages provided context relating to the buttons. The full description of the modalities used and their configuration is provided in Section 3.2.3.

Notification Timing

Three notifications were delivered in each game of Concentration. As each game was 60 seconds long (see Section 3.2.1), three notifications was determined to be the maximum number of notifications that could be delivered. In part, this was due to the long delivery times of the olfactory notifications, an issue which is discussed in Section 3.2.3. It was important that the 60 seconds could be automatically divided in a way that provided enough time for normal game behaviour as well as notification delivery. The variance in the primary task's pilot study (see Table 3.1) showed that some participants actually completed the game in less than a minute, so a buffer would be needed at the end of a game to ensure that all three notifications were delivered during the game. In addition, as multiple cards are removed from the game, it's likely that the mental demand of the task would decrease, finally dropping off significantly as the participant reaches the end of the game. A buffer was also needed at the start of the game to ensure that participants had enough time to build up a mental workload by playing the game uninterrupted.

The time in the middle of the game was divided up into 'timeslices', each of which was separated by 1 second. Notifications must start and end delivery within the timeslice, so a buffer was added at the end of each timeslice that equalled the time taken to deliver a notification. This meant that a notification could be delivered at any point in the timeslice and would have time to be delivered without overlapping with another timeslice. This is shown in Figure 3.4.

Splitting up the 60-second game in this manner allowed notifications to be delivered randomly during set 'windows' in the game, ensuring that all three could be delivered without overlapping. It also ensured that notifications would not be delivered too close to the end of the game, where a smaller number of cards might produce a smaller workload. The random element also helped to stop participants predicting when the next notification would occur.

Summary of Secondary Task

Isolating the secondary task from the experiment as a whole, the procedure was as follows:

1. The participant was introduced to all three messages and told which button to press when they received (*i.e.* saw, heard, felt or smelt) the message. Participants would signal their understanding by pressing the correct button after being shown the message.

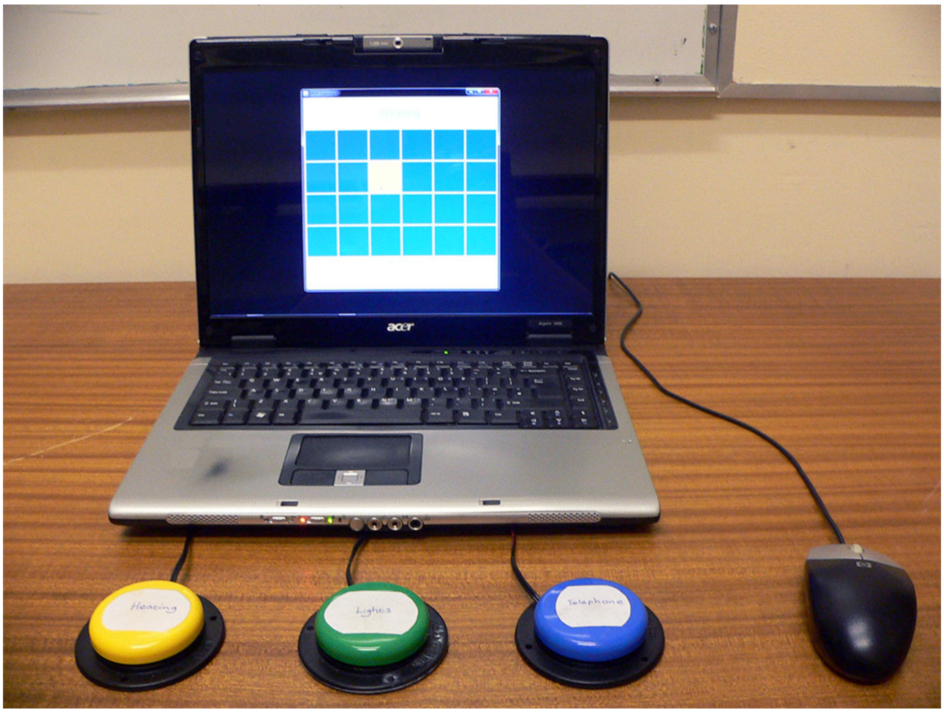


Figure 3.5: A photo showing the laptop and mouse used to carry out the primary task and the three buttons used for the secondary task. Participants were only required to use the three buttons and the mouse; they did not have to reach over the buttons to use the keyboard.

2. The participant would undertake a training regimen that delivered random notifications. Participants were expected to press the correct button to acknowledge the notification 6 times in a row to complete the training. The training would continue until they had accomplished this.

3. The participant would then start playing a game of Concentration (see Section 3.2.1). Three notifications were delivered during the game. When a notification was issued the participant must press the corresponding button. This step was repeated until all the games had been played.

The distribution of notifications over the games is described in Section 3.2.5. The standard setup of the primary and secondary tasks is shown in Figure 3.5.

3.2.3 Modality Selection & Notification Design

The aim of the study is to compare sensory channels, but as some sensory channels support several modalities, the study should consider a range of modalities to provide a reliable assessment of that sensory channel. Section 2.4.4 provided an exploration of several different modalities that can be used to deliver notifications. This section defines, describes and justifies the modalities that were tested during this study. The aims when selecting and configuring notifications were as follows:

1. each modality must be able to deliver at least 3 distinct messages;
2. the modalities should all be configured to deliver their payload in a similar amount of time;
3. the modalities should be configured in a way that reflects a real-world implementation of the modalities.

These aims were intended to ensure that a fair and ecologically valid comparison of the sensory channels. For all the modalities chosen the messages were delivered within 3 seconds. For modalities that could be continually delivered, such as text, they were delivered for the full 3 seconds; others such as speech stopped at the natural end point of delivery (*i.e.* they did not loop or repeat). After the 3 seconds, the participants had an additional 2 seconds to respond to the message before it was considered unacknowledged. The only exception to this was smell; the reasoning for this will be explained later in this section. Table 3.2 shows the final configuration of the modalities used in the study.

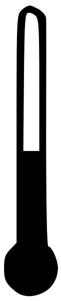
Text

Text is one of the most common ways to interact with a computer and uses the visual system (*i.e.* the human eye). The message itself can be changed simply by changing the text, however there are several presentation elements that can be manipulated to alter how the message is perceived, including the size, weight, colour, borders and animation [1, 117, 139]. As text is one of the most common modalities used in human-computer interaction, it was important to include it in the study.

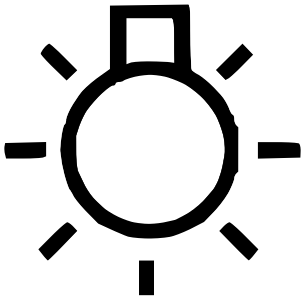
During the experiment textual messages were limited to a single word for each message: heating, lights and telephone. These terms were identical to the terms used on the buttons during the experiment. The text notification was displayed using a black, bold sans-serif font (specifically, 17pt DejaVu Sans) on a light-grey background. The message was not animated and would simply appear at the top of the play area. This configuration was chosen to ensure that the messages were legible and could be seen during play but without being overly-salient. As this message could be delivered indefinitely, *i.e.* it had no natural 'end of delivery' point, the message was displayed for 3 seconds.

Table 3.2: The configuration of the modalities used in the groundwork study of notification modality.

Modality	Message			Abstraction	Delivery Timing	Notes
	Heating	Lights	Telephone			
Text	"Heating"	"Lights"	"Telephone"	Explicit	Continuous	Delivered at the top of the game area.
Pictogram	ISO-7000 [107] Thermometer	ISO-7000 [107] Light	ISO-7001 [106] Telephone	Symbolic	Continuous	Delivered at the top of the game area.
Abstract Visual	Yellow #FFFF00	Green #008000	Blue #0000FF	Symbolic	Continuous	Coloured light delivered into peripheral vision.
Speech	Spoken "Heating"	Spoken "Lights"	Spoken "Phone"	Explicit	Finite	Delivered via audio, voice was synthetic 'Heather' from the Scottish Voice.
Auditory Icon	Gas Ignition	Light Switch Click (x2)	Dialling Beeps	Symbolic	Finite	Taken from a royalty-free sound-effects archive.
Earcon	Acoustic Grand Piano (Gen. MIDI #000)	Clarinet (Gen. MIDI #071)	Marimba (Gen. MIDI #012)	Abstract	Finite	From McGee-Lennon *et al.* [146], earcons varied in timbre.
Tacton	multiLP	textLP	voiceLP	Abstract	Finite	From Brown *et al.* [34], tactons varied in rhythm and roughness.
Olfactory	'Dark Chocolate'	'Riverside'	'Raspberry'	Abstract	Delayed Continuous	Delivered using Vortex device. Smell names defined by the Dale Air catalogue.



(a) A thermometer from ISO-7000 [107] was chosen to represent the 'heating' message.



(b) A light bulb from ISO-7000 [107] was chosen to represent the 'lights' message.

(c) A telephone handset from ISO-7001 [106] was chosen to represent the 'telephone' message.

Figure 3.6: The 3 pictograms used in the experiment to represent heating, lights and telephone were selected from various industry standard pictogram sets.

Pictogram

Along with text, pictograms are extremely common in HCI. They also use the visual system for interaction, however it is much harder to modify the message from a pictogram when compared to text. The interpretation of a pictogram is somewhat subjective, while text messages are explicit; this is one of the weaknesses of using pictograms. As pictograms are also common in HCI, pictograms were included in the experiment.

To match the text messages, the selected pictograms were black and displayed in the same area as the text. The pictograms were of a similar size to the textual messages. Selecting pictograms to represent the three messages proved difficult, but suitable pictograms were selected from two international standards documents: ISO-7000 "symbols for use on equipment" [107] and ISO-7001 "public information symbols" [106], as shown in Figure 3.6. The difficulty in finding pictograms was one of the reasons that the 'fan' message was replaced with 'telephone' (see Section 3.2.2); a suitable pictogram for 'fan' could not be found which could not also be interpreted to mean 'heating'. Like text, the pictograms were displayed for 3 seconds.

Abstract Visual

Unlike text and pictograms, abstract visual notifications can be difficult to interpret. A simple ambient message was chosen to represent abstract visual messages, in this case shining a coloured light onto the wall in the participant's peripheral vision as shown in Figure 3.7. Like the other visual methods this interacts with the visual system, but in this the interactions initially take place through peripheral vision.

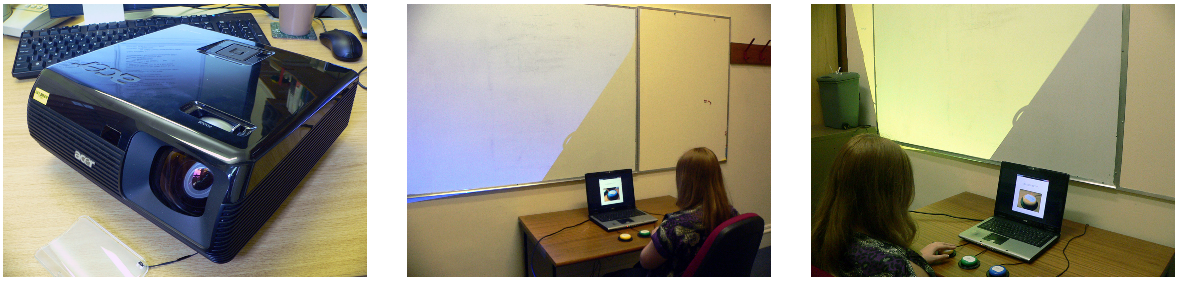


Figure 3.7: The abstract visual condition used a short-throw projector to create a coloured light in the peripheral vision of the participant. The blue and yellow lights are shown here to demonstrate the difference in intensity and contrast between the lights; a green light was also included but is not shown here.

Figure 3.8: Sennheiser HD 25-1-II headphones used during all audio conditions.

As one of the stated aims when selecting modalities was to make them as fairly comparable as possible, the colours chosen to represent each of the three messages matched the button to press in response, *e.g.* a red light signalled that the red button should be pressed. The colour of the light corresponded to the colour of the button to press. Like the other visual cues, the light was shown for 3 seconds.

Speech

Speech has been advocated by a wide range of researchers as an excellent interaction modality for older users [73, 145, 168, 224, 229]. Much like the text modality, speech messages can be changed by changing the message itself, but there are also several ways to alter the way the message is perceived (*e.g.*, how urgent the message is) [86]. While speech primarily uses the aural system (*i.e.* the human ears) the interpretation of a speech message can be influenced by visual stimuli [149].

As this experiment is based on the one carried out by McGee-Lennon *et al.* [146], the same synthetic speech engine[4] was used to create the speech messages for this study. The 'heating' and 'lights' messages were identical to those used by McGee-Lennon *et al.* while a new message was generated for 'telephone'. The speech messages, along with all other audio messages, were delivered through Sennheiser[5] HD 25-1-II headphones as shown in Figure 3.8. The headphones included padded cups to help prevent background noise from influencing the audio conditions, but did not use any form of active noise-cancellation.

As with text, the spoken messages were identical to the words written on the buttons; only a single word was delivered for each message and each message took around 1-1.5 seconds to deliver. Speech notifications were not repeated or looped. Participants were also free to select a volume that they found comfortable during the experiment, which helped to ensure that any subjective feedback was not affected by the audio being too loud or quiet.

Auditory Icon

The differences between speech and auditory icons are similar to the differences between text and pictograms. The principal shortcomings of pictograms are amplified for auditory icons; there is more room for interpretation and the auditory icons are more difficult to generate [67, 68]. As McGee-Lennon *et al.* [146] did not evaluate auditory icons in their experiment, auditory icons had to be selected that matched the other auditory notifications used in the experiment. The auditory icons used were taken from an online royalty-free sound effects archive,[6] as follows:

Heating A gas fire or boiler igniting.

Lights A light-switch being flicked on and off.

Telephone The sound of a telephone number being dialled.

The auditory icons were between 1 and 1.5 seconds long and were of a similar volume. The difficulty in finding an auditory icon to represent 'fan' was another reason why the 'fan' button was replaced with 'telephone'; as noted by Gaver [67], the primary

[4] The 'Heather' voice from The Scottish Voice, http://www.thescottishvoice.org.uk/.
[5] Sennheiser electronic GmbH & Co. KG, http://en-de.sennheiser.com/
[6] http://www.freesound.org/.

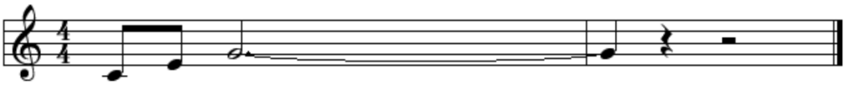


Figure 3.9: The three-note earcons used in the experiment. From McGee-Lennon *et al.* [146, p. 440].

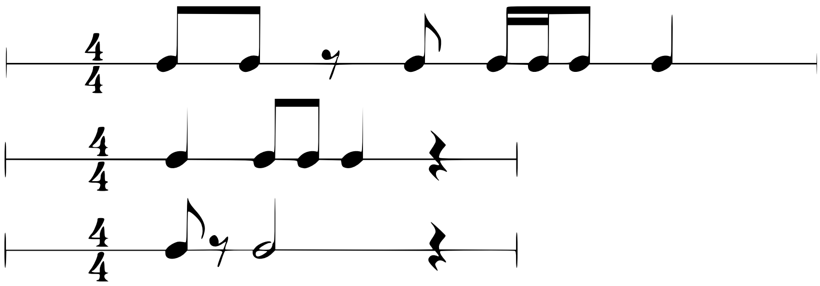


Figure 3.10: The three tacton rhythms used in the study. The tactons also varied in roughness (amplitude). From Brown *et al.* [34, p. 174].

issue auditory icons lies in finding suitable sounds. Like speech, auditory icons were not looped and were delivered through the headphones shown in Figure 3.8.

Earcon

Earcons are non-speech structured audio. There are several ways to configure earcons, including timbre, register, pitch, rhythm and intensity [32]. Earcons can be generated by computers using these variables, however these factors can also influence the way that earcons are interpreted, *e.g.* they can affect the perceived urgency [32, 58].

Like the speech messages, the earcons used were the same ones used by McGee-Lennon *et al.* [146]. Figure 3.9 shows the melody that was used for the earcons. Each earcon played the same rhythm but used 3 different timbres (instruments): a harpsichord, a clarinet and a marimba. As with speech and auditory icons, the earcons were not looped and delivered through the headphones shown in Figure 3.8.

Tactile

Tactile feedback is interesting for several reasons. Tactile-equipped mobile phones are effectively ubiquitous and tactile impairments are rare, as discussed in Section 2.1.1. Therefore, tactile feedback might be useful when delivering messages to impaired users. A structured tactile message (as opposed to a simple vibration) is called a Tacton [30],

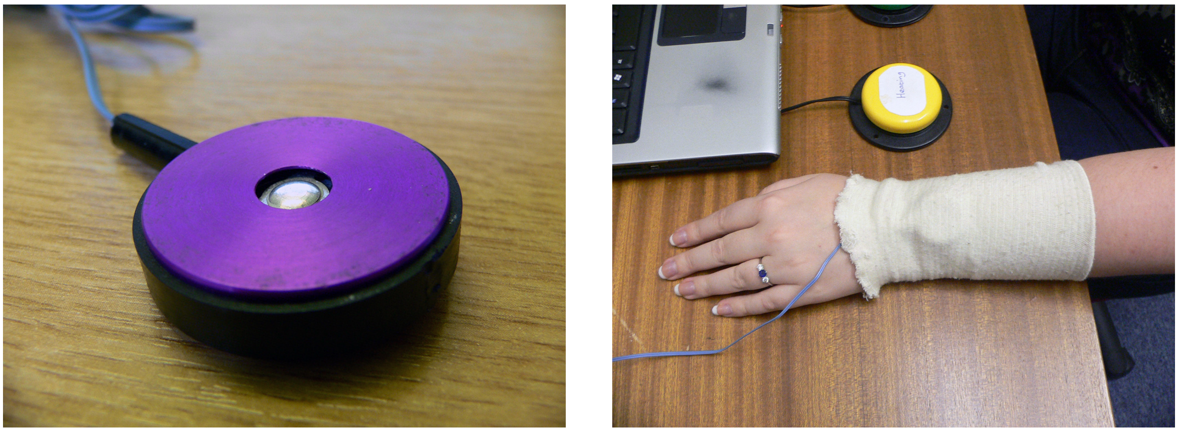


Figure 3.11: The tactile device used in the experiment was a C2, which was held against the wrist/forearm with a stretchable bandage.

Figure 3.12: The olfactory device used in the experiment. This figure shows how the fabric disks were inserted into the device and how the device itself was positioned close to the user to minimise delivery time.

and they can be generated by varying frequency, timbre, amplitude, waveform, duration and rhythm.

The tactons used in the experiment were taken from an experiment by Brown *et al.* [34]. They were delivered through an Engineering Acoustics[7] C2 Tactor shown in Figure 3.11. The C2 was placed on the upper-wrist of the participant's non-dominant hand, normally where a watch would be worn. This position was shown by McAdam & Brewster [136] as a good location for tactile feedback. The tactons themselves varied in rhythm and roughness (amplitude). The three Rhythms used are shown in Figure 3.10. Tactons were not looped and could be delivered in 1-2 seconds. Participants were able to adjust the intensity of the tactons to a comfortable level.

Olfactory

Olfaction uses the smell receptors in the human nose. Olfaction research is rare in HCI, generally because of the practical issues that are involved when trying to work with it. For example, Brewster *et al.* [31] reported a number of problems when working with smells, such as difficulty containing them when they were not wanted. Another issue was getting rid of smells after delivery. There may also be a very strong gender divide with smell interaction; Bodnar *et al.* [25] found that women were 51% more likely to wake up at the smell of smoke than men. However, smell devices do exist and some baseline smell research has been carried out.

The smell device used in the experiment was a Dale Air[8] Activ 5000X Vortex. The Vortex has 4 small fans which blow air over fabric discs infused with scent-providing chemicals as shown in Figure 3.12. The device has no method of preventing smells from dissipating into the air when the fans are stopped, and no way to clear smells from the air post-delivery. Brewster *et al.* [31] also used the Vortex device in their study.

To investigate which smells should be used, a range of smell discs were purchased including many based on the recommendations made by Brewster *et al.* [31]. After informally piloting the scents, the three disks chosen for the experiment were dark chocolate, riverbank (which smelled similar to eucalyptus) and raspberries. Dale air also provided an experimental disk which they claimed may be able to nullify a smell currently in the air; this disk was stored in the fourth chamber of the Vortex device. This 'smell-remover' disk appeared to have a small effect on the smells used but the experiment did not verify its effectiveness and Dale Air do not offer it as part of their standard catalogue.

Several concessions were made to the experimental design to allow the smell device to be used. One issue was the long length of time that smells took to deliver. In a pilot, the smell device took up to 12 seconds to deliver a message. For the experiment this was reduced by positioning the smell device close to the participant, as shown in Figure 3.12. The extended delivery times necessitated a longer response window for smell-based reminders; from the start of an olfactory message, participants had 15 seconds to respond (compared to 5 seconds for all the other modalities). The main potential adverse effect from this was an increased possibility that participants would predict when the next reminder was due. Participants in the experiment were warned

[7] Engineering Acoustics, `http://www.eaiinfo.com/`.
[8] Dale Air, `http://www.daleair.com/`.

that smells would linger after they pressed the button, but also that the same smell could be delivered twice in a row. Participants were asked to use their own judgement to decide if a smell was lingering or being delivered anew. Each smell was delivered for 10 seconds, with an additional 5 second window to respond.

When the smell disks were not in the Vortex device they were stored in individual plastic cases wrapped in foil and then plastic. The four disks were stored in separate Tupperware containers to prevent cross-contamination. When the smell condition was completed, the disks were returned to storage. The room was aired between experiments (and sometimes after the training segment if necessary) to prevent participants becoming desensitised to the smells.

The Vortex device used four small ebm-papst 255M[9] fans to deliver scents, which created a barely audible hum at < 5dB[10]. To prevent participants from using the sound of the fans as a timing cue for when to 'watch out' for smells, at least 1 of the fans was run at all times. When no smell was being delivered the device would switch to the 'smell nullifying' disk. The small fans were able to slow down and speed up extremely quickly, such that one fan could switched of and another on extremely quickly. The speed of this change and the low volume of the fans meant that any audio indicators of switching were minimal and posed little risk to the experiment (especially when compared to the risks posed by lingering smells).

3.2.4 Independent Variables

The independent variable for this study is the sensory channel used to receive the notifications, *i.e.* the visual, audio, tactile or olfactory systems. The 8 modalities used in the experiment consisted of 3 visual modalities, 3 aural modalities, 1 tactile modality and 1 olfactory modality. Each participant would complete 4 experimental conditions: 1 visual, 1 aural, 1 tactile and 1 olfactory. The visual and audio conditions counter-balanced the visual and audio modalities between participants to ensure an even coverage of the modalities. This meant that the number of participants in the experiment had to be a multiple of 3 to ensure even coverage of the modalities.

This suggests that the experiment is a mixed-methods design. However, the overall aim of this study was to provide an overview of the different ways in which different sensory

[9]Datasheer for ebm-papst 255M, http://img.ebmpapst.com/products/datasheets/DC-axial-fan-255M-ENG.pdf.

[10]The American Tinnitus Association lists a ticking watch at 20dB and a quiet whisper at 30dB. Source: http://www.ata.org/for-patients/how-loud-too-loud.

Table 3.3: This table shows how notifications were balanced over several games. In this case, three different messages are delivered over 3 games. Over the condition as a whole, each message is delivered three times.

	Notification 1	Notification 2	Notification 3
Game 1	Heating	Telephone	Telephone
Game 2	Lights	Telephone	Heating
Game 3	Lights	Heating	Lights

channels could be utilized for notification delivery. The 3 different modalities that make up the audio and visual conditions were intended to provide a balanced overview of the options for interaction in that sensory channel. Note that this study was followed by a similar study which examined each modality individually: that study is presented in Chapter 4 of this thesis.

Each participant would undertake 4 experimental conditions (visual, audio, tactile and olfactory) and a control condition without notifications. Each condition was composed of five games which were played sequentially for a total of 25 games: 20 with a secondary task (notifications) and 5 without. At the start of each condition a pseudo-random schedule was generated which ensured that all the notifications were delivered in equal amounts but in a random order over all the games. For example, if a single condition had three games with three notifications per game, then the schedule generated might look like the one shown in Table 3.3. This results in an ordering which appears random to the participants but ensures that the notifications themselves were counter-balanced.

3.2.5 Dependent Variables

The four research questions defined in Section 3.1 consider disruption, effectiveness, subjective assessment and MLP. This section will explore each of these terms and identify the measurements which can be taken to address the research questions.

Disruption

Disruption can manifest as a drop in activity rate or performance, and as such can be quite difficult to measure reliably. As discussed in Section 3.2.4, the experiment includes a control condition where no reminders were delivered. This control condition allows disruption to be measured by comparing the performance of the players between conditions which had notifications and the condition which did not. Therefore, the best way to measure disruption is to measure performance in the game.

Existing work concerning the performance of Concentration has used various techniques to measure performance. Most of these experiments were carried out through observation, and as such only a small amount of data could be recorded. As this experiment employs computers to log player data, a large amount of observations can be made. Not only does this allow for the observation of metrics developed by other researchers, but measurements can be made which would require a computer to gather effectively.

The number of cards matched or games completed are effective base measures of performance. However, they suffer from some important shortcomings. Firstly, they do not account for luck. Secondly, they may not accurately reflect disruption, as existing work shows that the presence of notifications will result in higher speed in ongoing tasks [115], and that the higher workload might produce higher performance [189]. In addition, cards matched and games won do not take a player's speed into account; a slower player is likely to match fewer cards and win less games, but it does not mean that player was more disrupted. *Cards matched per turn* could be used instead as a basic performance measurement, but this may not accurately reflect disruption.

Superfluous views is a metric that was devised specifically for the experiment. When a card has been viewed, that card is marked as 'seen'. Every subsequent viewing of that card which fails to match it to another card is considered a superfluous view. A high number of superfluous views suggests that the participant is finding it difficult to remember which cards are where. Compared to basic performance measures, superfluous views is of greater value as it represents a 'recovery behaviour' which should, theoretically, suggest greater disruption. It can be made independent of time by considering the *superfluous view per turn (SVpT).*

Baker *et al.* [13] suggested perfect pairs as a metric. A perfect pair occurs when a player clicks on a card for the first time and remembers having seen its pair; if the player successfully matches the first card to the second, it demonstrates a good awareness of where the cards are located and an efficient strategy. Perfect pairs is in a way the inverse of superfluous views. However, the pilot study described in Section 3.2.1 revealed that most participants will only make a very small number of perfect pairs; as such this measurement does not provide the same level of depth as SVpT.

Schumann-Hengsteler [181] proposed the idea of *omitted hits*[11]. An omitted hit is, in effect, a wasted turn. If the player has viewed two matching cards but fails to

[11] Schumann-Hengsteler credits the idea of 'omitted hits' to a poster by Baeriswl, Baeriswl-Rouiller & Etienne-Waeber.

match them, it is an omitted hit. Omitted hits would help to show disruption as it demonstrates an inability to remember the position of cards. However, the pilot revealed that the number of omitted hits will rise extremely quickly as a part of normal play, which is likely to result in highly skewed and unreliable data.

Gellatly *et al.* [69] observed strategy as a performance measurement as part of their work into investigating why older players outperformed younger played at Concentration. The two main strategies identified by Gellatly *et al.* were *recency* and *primacy*; both are used to match two cards that the player has already seen. With recency, the player first picks the most recently viewed card and then tries to match it. Conversely, the primacy strategy attempts to compensate for memory failures by looking for the first-viewed card first. If successful, the player can easily match to the most recent card; if not, then the player can still use the turn to attempt another match.

While these two strategies are excellent for searching for cards, there are two other scenarios that could lead to a pair of matched cards. The first is luck, where the player matches two cards where at least 1 of them had not been previously viewed. The other scenario is a perfect pair; this is *not* a recency strategy as the player is not making a decision about which card to select first. Therefore, *every* pair of cards matched can be classed as either a primacy match, recency match, perfect pair or lucky match.

While the strategy used during the game can tell us a lot about the player's skill (or possibly their strategic nous), it is unable to provide useful information about disruption. Gellatly *et al.* [69] noted that most players used the recency strategy, despite it being the least effective, and this is unlikely to change due to the presence of notifications. However, monitoring matches for perfect pairs and for lucky matches could be useful in interpreting the results, in particular when differentiating between a lucky player and one with excellent memory.

Therefore, the primary metrics that will used to measure disruption are as follows:

Matches per Turn Matches per turn is a simple measure of game performance independent of game completion or player speed. It is calculated by taking the number of matches per turn and dividing it by the number of turns made in a game. This measurement can range from 0 to 1.[12]

[12] A score of 0 would mean no matches were made. A score of 1 would mean a match was made every turn, which is only possible through luck.

Superfluous Views per Turn (SVpT) SVpT measures the frequency of a disruption-recovery behaviour (superfluous views) as the game goes on. It is calculated by taking the number of superfluous views and dividing it by the number of turns made in a game. This measurement can range from 0 to < 2.[13]

Effectiveness

Notification effectiveness was measured in two ways. Firstly, the time taken for the message to be delivered to the participant was measured as the *notification response time*, or simply response time. This measurement represents the time in seconds between the initiation of notification delivery and the participant pressing a button in response. If the participant failed to respond to a notification then this was marked as missing data. For each condition the average response time was taken.

The other effectiveness measurement was the accuracy of responses to the notification. Accuracy was measured by classifying the three possible outcomes of a notification; the correct button was pressed, the wrong button was pressed, or the notification went unacknowledged. For simplicity, this was represented as a percentage. This gave the following measurements for effectiveness:

Average Response Time (seconds) The time in seconds between a notification starting delivery and the participant responding to it (or the response time-out occurring).

Percentage of Responses Correct, Incorrect & Missed This is a simple measurement of the number of responses that were correctly acknowledged, incorrectly acknowledged and not acknowledged. This is expressed as a percentage for ease of interpretation; each of the three values could therefore range from 0 to 1.

Subjective Assessment

The primary subjective assessment was carried out using the NASA Task-Load Index (NASA-TLX), originally developed by Hart & Staveland [79]. The original NASA-TLX consisted of 2 parts:

[13] A score of 0 would be a perfect game, which can be done without luck but would require an excellent memory. It is actually impossible to reach 2; the worst-case scenario is $\left[2 - \left(\frac{2}{t}\right)\right]$, where t is the number of turns in a given game. This accounts for the first turn of the game.

1. A survey consisting of 6 scales;
2. A pair-wise comparison of those six scales, which would produce a weighting that could be applied for improved accuracy.

Twenty years after the development of NASA-TLX, Hart [78] evaluated the way that it had been applied in research. Hart found that the weighting process was often omitted without having a significant effect on the reliability of the survey. To maximise the time available for the experimental conditions, only the first part of the NASA-TLX was used in the study (known as Raw TLX).

NASA-TLX provides seven measurements through six 21-point Likert scales. The six scales in the survey were defined by Hart & Staveland [79, p. 169] as follows:

Mental Demand (MD) – How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?

Physical Demand (PD) – How much physical activity was required? Was the task easy or demanding, slack or strenuous?

Temporal Demand (TD) – How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?

Overall Performance (OP) – How successful were you in performing the task? How satisfied were you with your performance?

Frustration (FR) – How irritated, stresses, and annoyed versus content, relaxed, and complacent did you feel during the task?

Effort (EF) – How hard did you have to work (mentally and physically) to accomplish your level of performance?

The final value given by NASA-TLX is Workload (WL). Workload is calculated by the sum of the 6 scales, and represents the overall cognitive load of an exercise.

NASA-TLX is useful for several reasons: without the weighting component it can be quickly administered and it is a robust and well-known assessment tool [78]. The most useful function of NASA-TLX is that each of the 6 scales can be evaluated individually, which will help in understanding the properties of the different modalities. It will also allow a comparison between perceived and actual performance. The TLX form used in the study is provided in Appendix A.

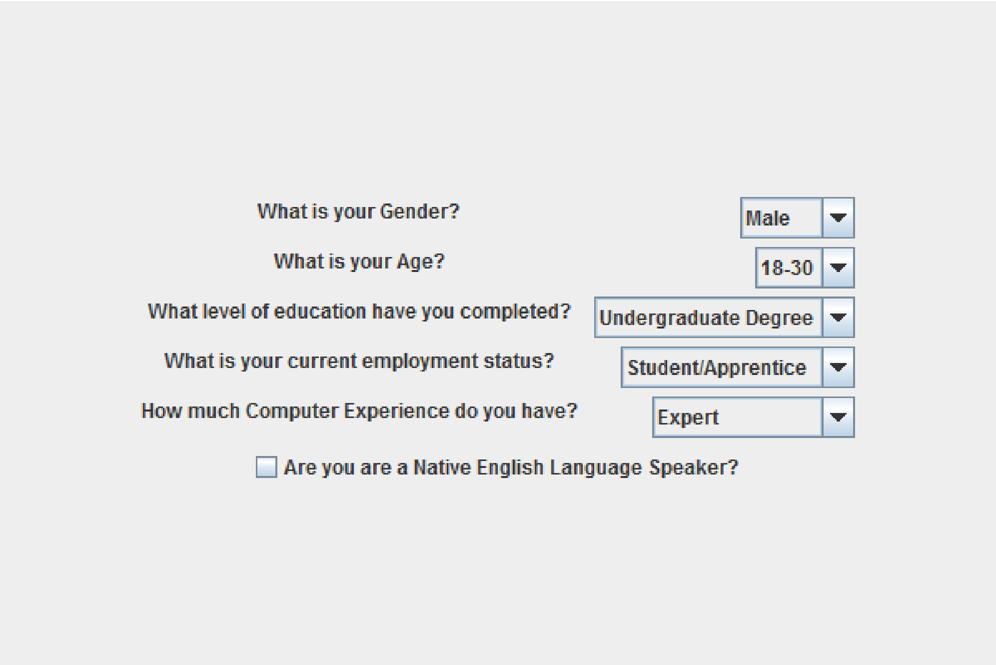
The second way in which informal data was gathered was through an informal exit interview. This was recorded by taking shorthand notes on the back of participant's NASA-TLX forms, and was intended to provide further insights into the participants' opinions of the different modalities.

Modal Learning Preference

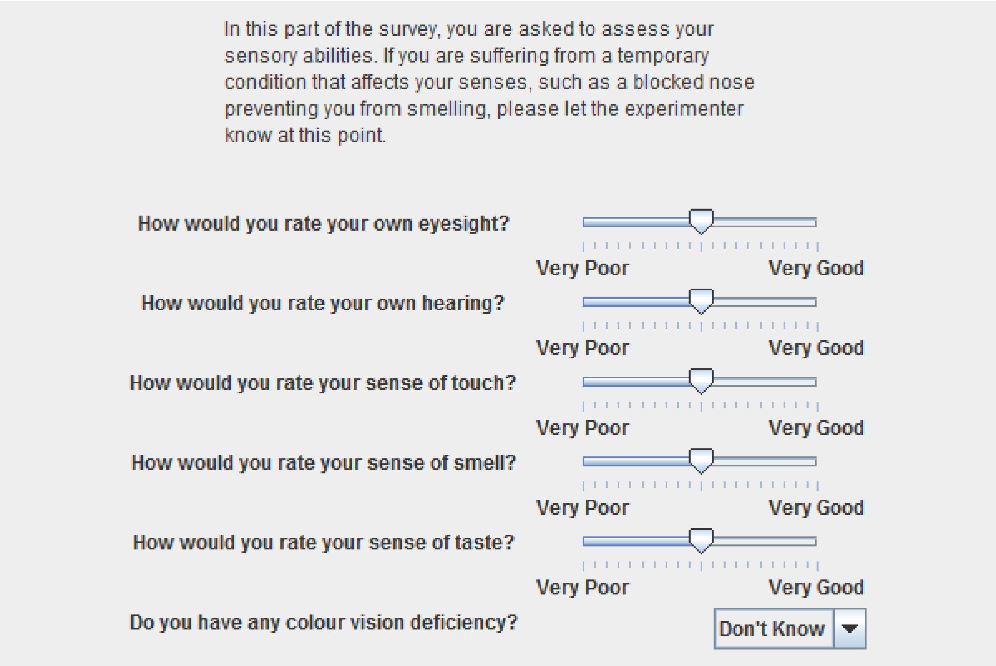
MLP theory holds that people will have a natural 'preferred' way of learning, *e.g.* by reading or by doing. The preferred way of learning is called the Modal Learning Preference (MLP). MLP theory does not suggest that people will have a *single* preferred modality; each modality is effectively independent, and having a higher score in one will not affect your score in another [65]. MLP is typically assessed through a simple paper-based questionnaire and the data mainly used to personalise teaching methods. However, this practice is controversial as some researchers claim there is a lack of reliable evidence to support this application [170, 184]. While critics such as Pashler *et al.* [170] take issue with using MLP to guide educational practices, Pashler *et al.* also stated that "*the existence of preferences with some coherence and stability is not in dispute*" [170, p. 108].

There are several schools of though on the taxonomy of modes that can be used for learning. The most popular are VAK (Visual, Audio and Kinaesthetic) and VARK (Visual, Audio, Reading/Writing and Kinaesthetic), both of which are based on the work of Fleming & Mills [65]. As the experiment has visual, auditory and tactile conditions (see Section 3.2.4), the VAK model provides the most direct mapping of learning preferences to experimental conditions.

The VAK survey used was a 30-question survey sourced from Brookhaven College[14] and was based on the Barsch Learning Style Inventory [15]. This survey produced three scores for each of the visual, auditory and kinaesthetic scales that can range from 10 to 30. For simplicity these values were normalised to range from 0 to 20.



(a) The general demographic component of the survey.



(b) The sensory impairment component of the survey.

Figure 3.13: The survey administered to participants at the start of the experiment.

3.2.6 Confounding Variables

There are several potential confounding factors that could affect the study. As there are several examples of interference between information and 'noise' in a sensory channel [91, 123], any potential background noise was minimised. This was achieved by carrying out the experiment in the same location under normal office lighting in a quiet area of the building. Headphones were worn during audio conditions and the room was aired between participants.

As discussed in Section 2.1, sensory decline occurs naturally with age. Gender differences may also affect the results, in particular when considering smell [80]. To manage this effort was made to ensure participants were balanced between male and female, and that a wide range of ages were included. The experiment did not however attempt to balance participants over age groups.

Gender and age information was gathered as part of a demographic survey. The survey also asks users to provide information about their education, employment and computer experience. This data was used to ensure that the experimental participants represented a reasonable cross-section of the general population and avoid accusations of selection bias. The survey presented to participants is shown in Figure 3.13a.

Due to the impact it might have on the results, it was important that participants did not have severe sensory impairments. To evaluate this, a short self-assessment questionnaire was developed which asked participants to evaluate their own sensory ability on a 21-point Likert scale modelled on the NASA-TLX format. This is shown in Figure 3.13b.

3.3 Hypotheses

Section 3.2 presented the design of the experiment, which included in Section 3.2.5 the measurements that would be taken to evaluate the four research questions given in Section 3.1. This section refines the research questions into the hypotheses that were tested during the study.

[14] http://www.brookhavencollege.edu/learningstyle/

3.3.1 Research Question 1 – Disruption

Research question 1 asked "*does the sensory channel used to receive a notification affect how disruptive the notification is to an ongoing task?*". As discussed in Section 3.2.5, one way to measure disruption is to compare performance between interrupted and uninterrupted tasks. As research shows interruptions can sometimes lead to *improved* performance [115, 189], performance alone may not accurately reflect disruption. The behaviour of the player was also observed to reveal more subtle behavioural changes.

Research has shown that performance will normally drop when a task is interrupted [11], but existing work has not yet fully explored the relationship between disruption and sensory channel [9, 25, 146]. Existing work also notes that activity rate will change for an interrupted task [115]. It may be possible for the player to speed up and maintain the same level of performance [39, 134], so a measure of errors made in the primary task will provide a useful insight into disruption. Therefore the first three hypotheses are as follows:

H 1.1 The sensory channel used to receive a notification will affect performance in the card matching game *(measured as the Cards Matched (CM)*.

H 1.2 The sensory channel used to receive a notification will affect the activity rate in the card matching game *(measured in clicks per minute)*.

H 1.3 The sensory channel used to receive a notification will affect the error rate in the card matching game *(measured in superfluous clicks per turn)*.

3.3.2 Research Question 2 – Effectiveness

Research question 2 asked "*are some sensory channels more effective than others at receiving information?*". As discussed in Section 3.2.5, there are two primary performance measures: response speed and response accuracy. Given the inherent differences between the sensory channels (which is discussed in Section 3.2.3) it is expected that the different notifications will produce very different performance levels. This was tested with two hypotheses as follows:

H 2.1 The sensory channel to which notifications are delivered will affect the speed at which participants respond to notifications *(measured as the average response time).*

H 2.2 The sensory channel to which notifications are delivered will affect the ability to perceive and interpret notifications *(measured as the percentage of responses correct, incorrect & missed).*

There may also be a strong effect of gender in the olfactory condition: Bruck & Brennan [36] found that 80% of females would wake at the smell of smoke compared to 29% of males. The effects of gender on olfactory performance was tested by the following hypothesis:

H 2.3 Female participants will outperform male participants when responding to olfactory notifications *(measured by the response speed and the percentage of correct responses).*

3.3.3 Research Question 3 – Subjective Factors

Research question 3 asked "*what are the subjective assessments of the different sensory channels, and does perceived performance correlate to actual performance?*". Research has shown that the introduction of interruptions will increase stress, annoyance, anxiety and perceived difficulty [12, 46, 134]. Research has not revealed a relationship between such subjective factors and the sensory channel to which notifications are delivered. This was tested by the following hypothesis:

H 3.1 The sensory channel to which notifications are delivered will affect the subjective workload involved with the task *(measured by NASA-TLX).*

A lack of correlation between perceived performance and actual performance can indicate uncertainty or a loss of confidence. Due to the simplicity of the tasks, participants are expected to be able to assess their own performance without problems; however, this may be an issue in the conditions where notifications are abstract (*e.g.* tactile and olfactory). This was tested as follows:

H 3.2 The perceived performance *(measured by NASA-TLX Overall Performance (OP))* will not correlate to actual performance *(measured by Cards Matched (CM) and percentage of correct responses).*

3.3.4 Research Question 4 – Modal Learning Preference

Research question 4 asked "*does Modal Learning Preference (MLP) correlate to an improved ability to receive information in that sensory channel?*". The VAK survey described in Section 3.2.5 consists of three parts: visual, auditory and kinaesthetic. The visual and audio learning preferences are expected to match up to the visual and audio conditions, but kinaesthetic learning preference may not match the tactile condition. Kinaesthetic learning means 'learning by doing', which can be described as a repeated process of taking an action then learning from the reaction. While this heavily implies active tactile interaction, this may not use the same resources as passively receiving tactile information; as such no correlations are expected between kinaesthetic learning preference and tactile conditions. The following hypotheses were tested to address the research question:

H 4.1 A higher visual learning preference will correlate to improved reaction times to visual notifications.

H 4.2 A higher visual learning preference will correlate to improved response accuracy *(measured as the percentage of responses correct)* for visual notifications.

H 4.3 A higher audio learning preference will correlate to improved reaction times to audio notifications.

H 4.4 A higher audio learning preference will correlate to improved response accuracy *(measured as the percentage of responses correct)* for audio notifications.

H 4.5 Tactile and Olfactory reaction times will not correlate to any of the MLP measures.

H 4.6 Tactile and Olfactory response accuracy will not correlate to any of the MLP measures.

3.4 Participants

A call for participants was distributed to University students via an e-mail list. Participants were also sourced via Facebook and Twitter in an attempt to find more older participants. A total of 27 participants were included in the experiment of which 13 were male and 14 female. The participants were mostly students aged 18-30 ($n = 20$) but there were also participants aged 31-45 ($n = 4$) and 46-60 ($n = 3$). Although a good gender split was desired when sourcing participants, age was not controlled which resulted in an uneven distribution. There were no notable imbalances in the observed demographic measures (computer experience, education level and employment). None of the participants reported any significant sensory impairments.

3.5 Procedure

At the start of the experiment, the participant was given an information sheet and consent form, both of which are provided in Appendix A. Once the consent form had been signed, the participant was asked to fill out the demographic and sensory impairment survey shown in Figure 3.13.

The control condition was run first in each trial followed by the experimental conditions, which were delivered in a random order over all the participants. The type of visual and audio notifications that could be delivered were counter-balanced over the whole participant group, so 9 participants would see each of the visual and audio methods. As stated in Section 3.2.3, the study was concerned with the sensory channel used, and different modalities were used to provide a sample of the types of interactions available in that channel.

The experimental conditions started with a screen that explained the modality that would be used for interaction in that condition. At this point any equipment that was required (*e.g.* headphones, smell device, C2) was set up and adjusted with the participant to ensure comfort. The participant would then begin the training segment.

The training segment started by demonstrating the three notifications (corresponding to heating, lights and telephone as discussed in Sections 3.2.2 and 3.2.3) to the participant and asking them to press the correct button to proceed. Once completed, 6 notifications (2 for each notification) were delivered in a random order, and participants were expected to correctly respond to all 6. If they did so, they would proceed to the next stage of the experiment; if they made a mistake they would return to the start of the training. No training was required for the control condition.

Once the training was completed participants would play 5 games of concentration, with each lasting a maximum of 60 seconds as specified in Section 3.2.1. No notifications were delivered during the control condition, but during the experimental conditions 3 notifications were delivered per game as specified in Section 3.2.2. The notifications were counter-balanced over all 5 games, so the participants would always receive 5 of each notification.

With all 5 games complete participants were asked to complete a NASA-TLX survey on paper while the experimenter cleared away any equipment and prepared for the following condition. This process was repeated until all conditions were complete.

At the end of the experiment participants were asked to complete the VAK MLP survey described in Section 3.2.5 (this is provided in Appendix A). The participant was then paid and an informal interview was carried out, during which the interviewer took notes on the back of the VAK survey sheet. The experiment required 45-60 minutes and participants were paid £6 for their time.

3.6 Results

This section presents the results of the study, organised by the hypotheses laid out in Section 3.3.

3.6.1 Hypothesis 1.1 (Performance)

Hypothesis 1.1 was defined in Section 3.3.1 as "the sensory channel used to receive a notification will affect performance in the card matching game". This was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and Cards Matched (CM) as the dependent variable. No other factors were included in the model.

Mauchley's test was found to be significant ($\chi^2(9) = 29.33,\ p < .05$), indicating that the assumption of sphericity had been violated; this was addressed by applying the Greenhouse-Geiger correction ($\epsilon = .64$). The final model showed that the sensory channel had a significant main effect on the cards matched in a game ($F(2.56,\ 58.93) = 17.21,\ p < .001,\ \eta^2 = .43$). This result is shown in Figure 3.14.

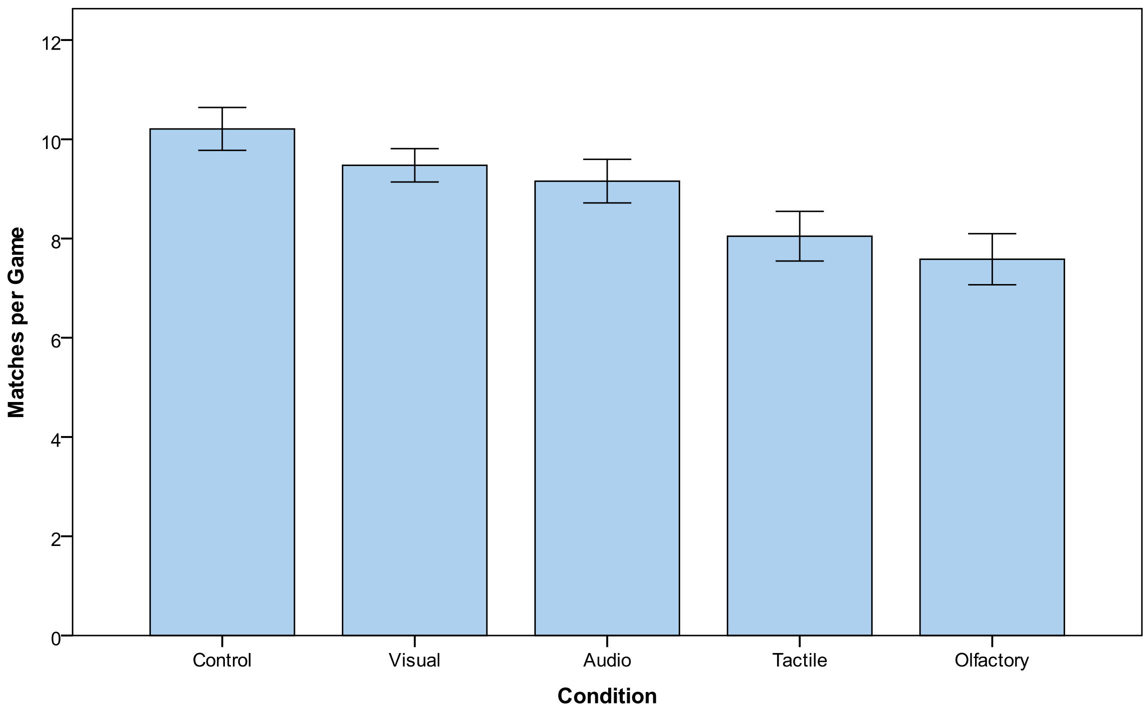


Figure 3.14: Graph showing the relationship between the condition and the mean number of cards matched in a game. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Table 3.4: Significance (p) values of *post hoc* pairwise comparisons of the condition on the mean number of cards matched.

	Con		Vis		Aud		Tac		Olf	
Con	-		.034	*	.027	*	.001	**	.000	***
Vis	.034	*	-		.681		.020	*	.000	***
Aud	.027	*	.681		-		.158		.025	*
Tac	.001	**	.020	*	.158		-		.482	
Olf	.000	***	.000	***	.025	*	.482		-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Post hoc pairwise comparisons were calculated using the estimated marginal means and corrected using the Sidak correction, as shown in Table 3.4. This showed that the number of cards matched dropped significantly between the control conditions and the conditions with notifications. Within the experimental conditions some conditions were not significantly different: visual and audio, audio and tactile, and tactile and olfactory.

The results demonstrate that the presence of notifications had a negative effect on the average number of matches made in a game. The results also show that the tactile and olfactory notifications appeared to slightly lower performance in the primary task. The results confirm the hypothesis: the sensory channel used to receive a notification affected performance in the card matching game.

3.6.2 Hypothesis 1.2 (Activity Rate)

Hypothesis 1.2 was defined in Section 3.3.1 as "the sensory channel used to receive a notification will affect the activity rate in the card matching game". This was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and turns per second as the dependent variable. No other factors were included in the model.

Mauchley's test was found to be significant ($\chi^2(9) = 19,\ p < .05$), indicating that the assumption of sphericity had been violated; this was addressed by applying the Greenhouse-Geiger correction ($\epsilon = .7$). The final model showed that the sensory channel had a significant main effect on the activity rate in a game ($F(2.8, 64.37) = 7.57,\ p < .001,\ \eta^2 = .25$). This result is shown in Figure 3.15.

Post hoc pairwise comparisons were calculated using the estimated marginal means and corrected using the Sidak correction. Table 3.5 shows the result of the pairwise comparisons, which revealed that the activity rate differences were primarily between the visual and olfactory conditions, although there was also a significant difference between the audio and tactile conditions.

Figure 3.15 shows that the activity rate rose slightly (but not significantly) in the visual and audio conditions, which was expected based on existing research. However for the tactile and olfactory conditions the mean activity rate fell below that of the control condition (again, not significantly). This suggests that participants sped up slightly when faced with the visual and olfactory tasks but the additional overheads of dealing

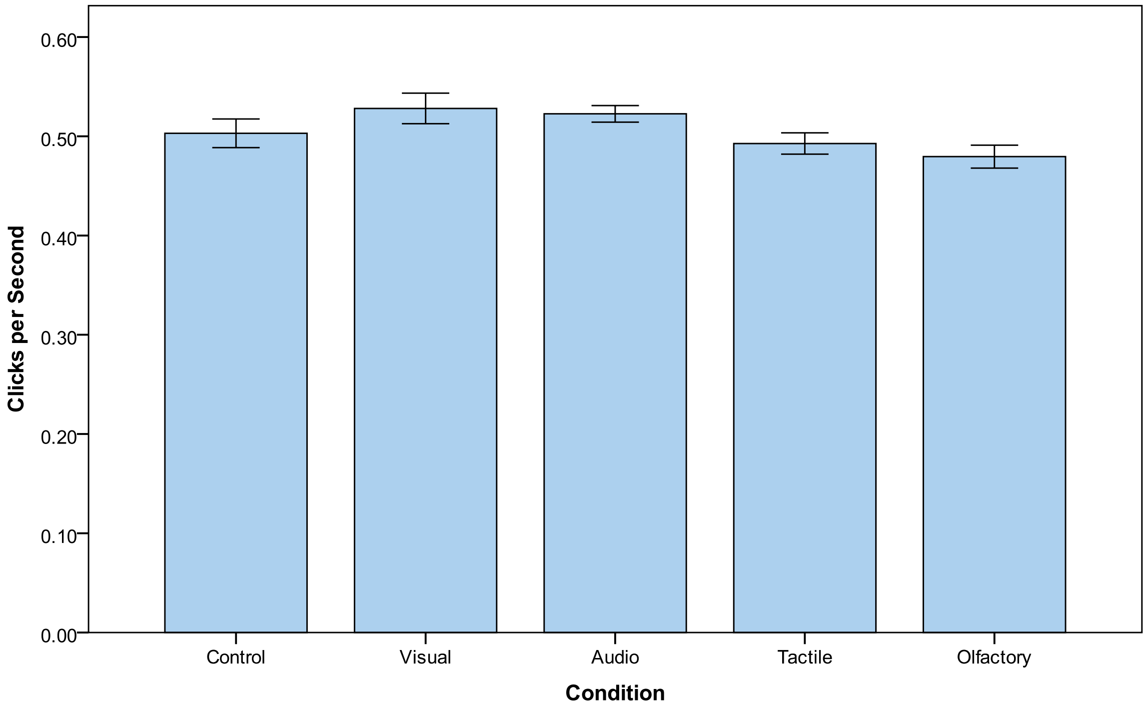


Figure 3.15: Graph showing the relationship between the condition and the mean activity rate (in clicks per second). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Table 3.5: Significance (p) values of *post hoc* pairwise comparisons of the condition on the mean activity rate (clicks per second).

	Con		Vis		Aud		Tac		Olf	
Con	-		.800		.831		.895		.116	
Vis	.800		-		.999		.140		.006	**
Aud	.831		.999		-		.035	*	.000	***
Tac	.895		.140		.035	*	-		.282	
Olf	.116		.006	**	.000	***	.282		-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

with olfactory and tactile notifications caused a reduction in activity rate in those conditions, leading to the significant differences. This partly confirms the hypothesis and will be discussed in more detail in Section 3.7.

3.6.3 Hypothesis 1.3 (Error Rate)

Hypothesis 1.3 was defined in Section 3.3.1 as “the sensory channel used to receive a notification will affect the error rate in the card matching game”. This was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and superfluous views per turn as the dependent variable. No other factors were included in the model.

Mauchley's test was not found to be significant ($\chi^2(9) = 2.29,\ p = .99$), indicating that the data satisfied the assumption of sphericity. The model showed that the sensory channel had a significant main effect on the error rate in a game ($F(4,\ 92) = 14.22,\ p < .001,\ \eta^2 = .38$). The results are shown in Figure 3.16.

Post hoc pairwise comparisons were calculated using the estimated marginal means and corrected using the Sidak method. The result are shown in Table 3.6, which revealed that there were significant differences between the control condition and the experimental conditions, but no other significant differences. The results do not support the hypothesis and instead seem to suggest that a similar level of disruption was produced regardless of the sensory channel used for notifications. This is discussed further in Section 3.7.

3.6.4 Hypothesis 2.1 (Response Time)

Hypothesis 2.1 was defined in Section 3.3.2 as “the sensory channel to which notifications are delivered will affect the speed at which participants respond to notifications”. This was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and response time as the dependent variable. No other factors were included in the model.

Mauchley's test was found to be significant ($\chi^2(5) = 46.84,\ p < .05$), indicating that the assumption of sphericity had been violated; this was addressed by applying the Greenhouse-Geiger correction ($\epsilon = .47$). The final model showed that the sensory channel had a significant main effect on the notification reaction time ($F(1.34,\ 32.17) = 285.25,\ p < .001,\ \eta^2 = .93$). This result is shown in Figure 3.17.

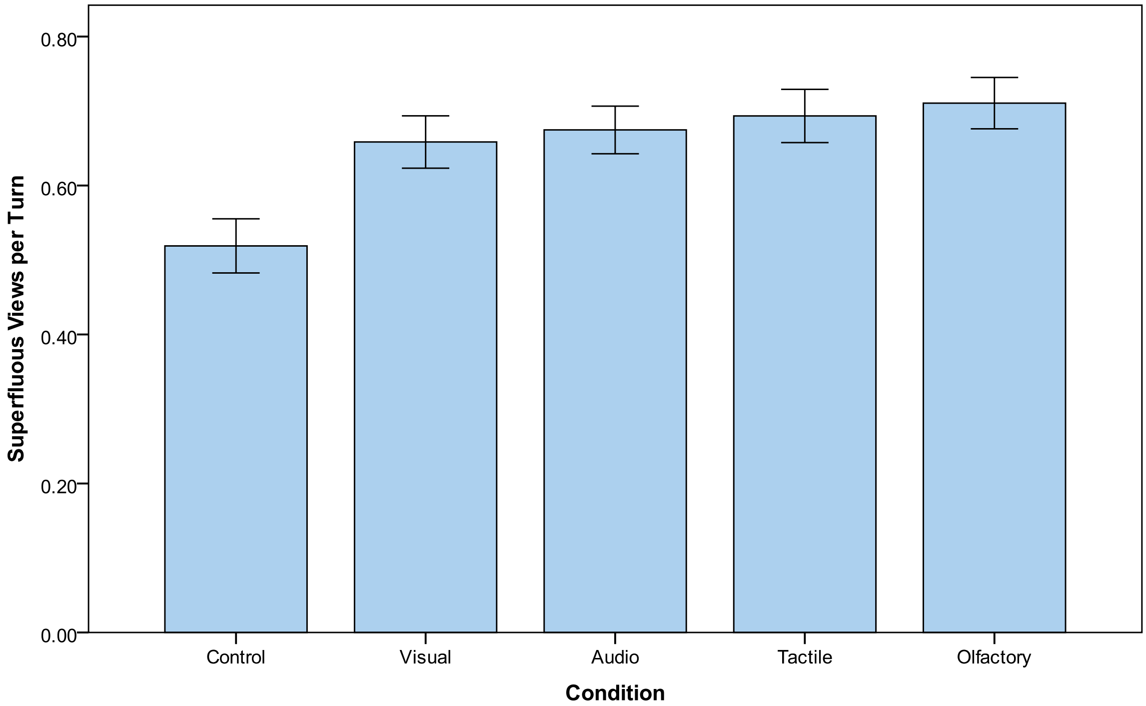


Figure 3.16: Graph showing the relationship between the condition and the mean error rate (superfluous views per turn). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Table 3.6: Significance (p) values of *post hoc* pairwise comparisons of the condition on the mean error rate (superfluous views per turn).

	Con		Vis		Aud		Tac		Olf	
Con	-		.001	***	.000	***	.000	***	.000	***
Vis	.001	***	-		.981		.897		.353	
Aud	.000	***	.981		-		1.000		.915	
Tac	.000	***	.897		1.000		-		.990	
Olf	.000	***	.353		.915		.990		-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

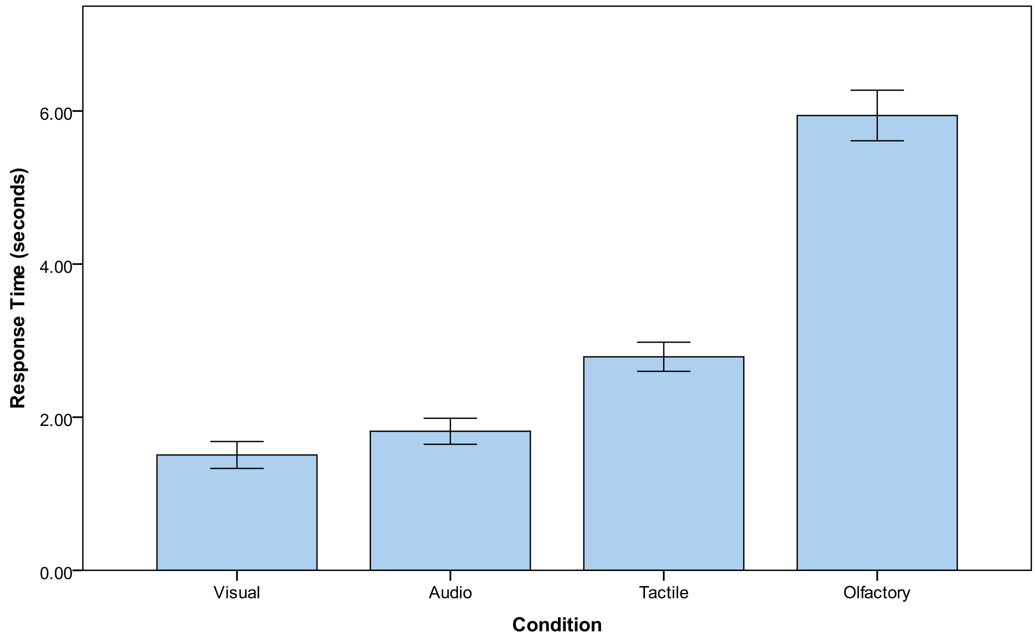


Figure 3.17: Graph showing the relationship between the condition and the mean reaction time. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Table 3.7: Significance (p) values of *post hoc* pairwise comparisons of the condition on the mean reaction time.

	Vis		Aud		Tac		Olf	
Vis	-		.005	**	.000	***	.000	***
Aud	.005	**	-		.000	***	.000	***
Tac	.000	***	.000	***	-		.000	***
Olf	.000	***	.000	***	.000	***	-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

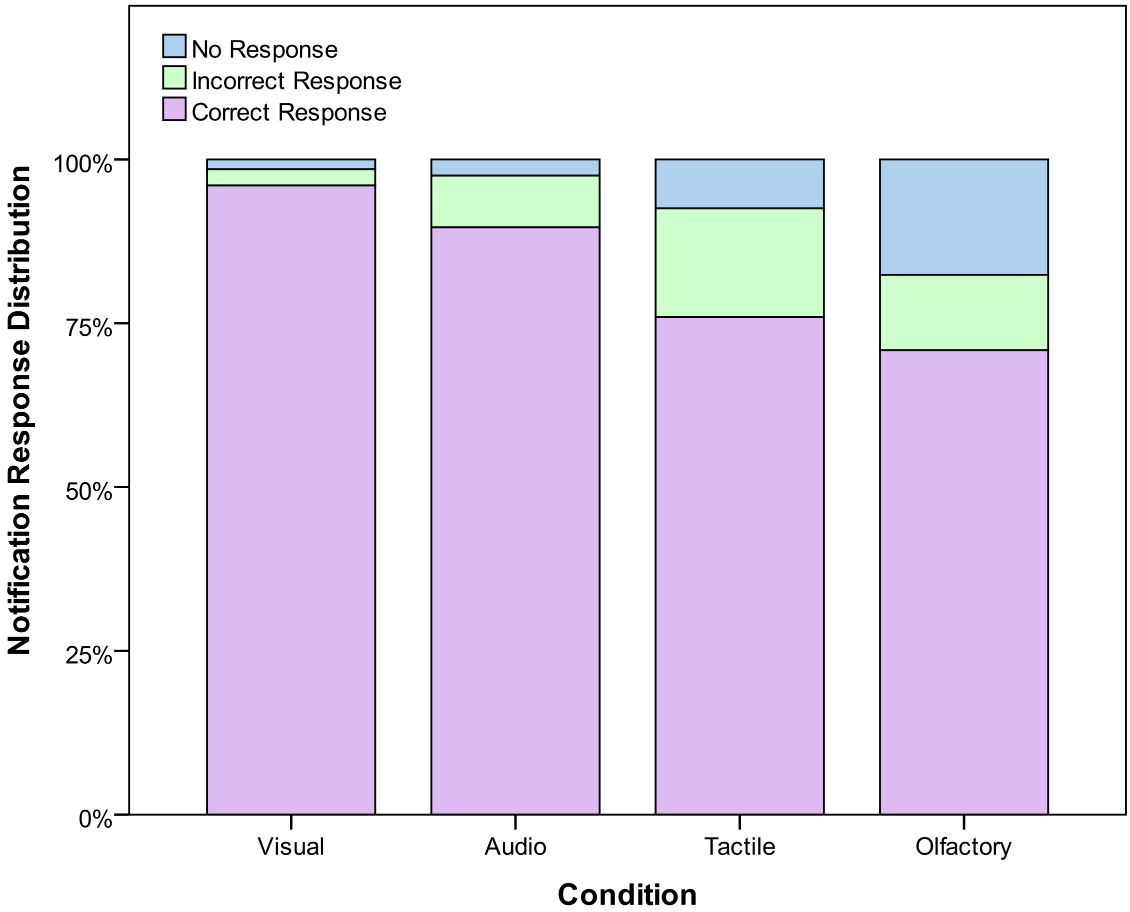


Figure 3.18: Graph showing the distribution of responses within each condition.

Post hoc pairwise comparisons were calculated using the estimated marginal means and corrected using the Sidak correction. The results, as shown in Table 3.7, reveal that all of the methods were significantly different from each other. Visual methods were the quickest, followed by the audio and then the tactile methods. The olfactory method took much longer than all the others to deliver the notification. This evidence supports the hypothesis that the sensory channel used had a significant effect on reaction time.

3.6.5 Hypothesis 2.2 (Response Accuracy)

Hypothesis 2.2 was defined in Section 3.3.2 as "the sensory channel to which notifications are delivered will affect the ability to perceive and interpret notifications". There were 3 variables to test; percentage of reminders correct, incorrect and unacknowledged. While these three factors are related, it was still necessary to create three general linear models.[15] The distribution of responses is shown in Figure 3.18.

[15] With two factors, only one GLM would be needed as the second factor would simply be the inverse of the first. With three factors, all three required evaluation.

Correct Responses

Number of correct responses was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and percentage of correct responses as the dependent variable. No other factors were included in the model.

Mauchley's test was found to be significant ($\chi^2(5) = 12.94,\ p < .05$), indicating that the assumption of sphericity had been violated; this was addressed by applying the Greenhouse-Geiger correction ($\epsilon = .74$). The final model showed that the sensory channel had a significant main effect on the correct responses in a game ($F(2.23,\ 51.27) = 10.42,\ p < .001, \eta^2 = .31$). This result is shown in Figure A.1a.

Post hoc pairwise comparisons were calculated using the estimated marginal means and corrected using the Sidak correction. This revealed that the tactile and olfactory conditions were significantly different from the visual conditions, but not from each other, suggesting two groups. There was a high amount of variance in the auditory condition and it was not significantly different from either group. These results are shown in Table A.1b.

Incorrect Responses

Number of incorrect responses was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and percentage of incorrect responses as the dependent variable. No other factors were included in the model.

Mauchley's test was found to be significant ($\chi^2(5) = 19.68,\ p < .05$), indicating that the assumption of sphericity had been violated; this was addressed by applying the Greenhouse-Geiger correction ($\epsilon = .7$). The final model showed that the sensory channel had a significant main effect on the incorrect responses in a game ($F(2.11,\ 48.57) = 3.85,\ p < .05,\ \eta^2 = .14$). This result is shown in Figure A.2a.

Post hoc pairwise comparisons were calculated using the estimated marginal means and corrected using the Sidak correction as shown in Table A.2b. This revealed the same grouping as correct responses: tactile and olfactory conditions were significantly different from the visual condition, while the audio condition was not significantly different from either group, mainly due to the higher variance.

Unacknowledged Responses

Number of correct responses was tested by constructing a repeated-measures general linear model (GLM) with sensory channel as the independent variable and percentage of correct responses as the dependent variable. No other factors were included in the model.

Mauchley's test was found to be significant ($\chi^2(5) = 25.64,\ p < .05$), indicating that the assumption of sphericity had been violated; this was addressed by applying the Greenhouse-Geiger correction ($\epsilon = .58$). The final model showed that the sensory channel had a significant main effect on the unacknowledged notifications in a game ($F(1.73,\ 39.74) = 11.73,\ p < .001,\ \eta^2 = .34$). This result is shown in Figure A.3a.

Post hoc pairwise comparisons were calculated using the estimated marginal means, and corrected using the Sidak correction. Table A.3b shows that the olfactory condition was significantly different from the visual and audio conditions, but none of the other conditions were significantly different.

Assessment

The results confirm the hypothesis that the sensory channel would affect notification response accuracy, however the different conditions were more similar than anticipated. The results revealed several interesting points: there was a high amount of variance in the audio conditions; the olfactory notifications were likely to be missed, but otherwise were quite effective; and that tactile notifications were less likely to be missed but were more likely to be incorrectly acknowledged. These results will be explored in Section 3.7.

3.6.6 Hypothesis 2.3 (Gender & Olfactory Performance)

Hypothesis 2.3 was defined in Section 3.3.2 as "female participants will outperform male participants when responding to olfactory notifications". This was tested by examining the response time and the percentage of correct responses. There was a difference between the mean response times of male ($M = 6.41,\ n = 12,\ \sigma = 1.57$) and female ($M = 5.65,\ n = 13,\ \sigma = .89$) participants; this was found to be significant by a one-tailed independent samples T test ($t(23) = -1.87,\ p < .05$). The results did not show an effect of gender on the percentage of correct responses ($t(23) = .94,\ p = .36$). The results partly confirm the hypothesis that female participants would perform better

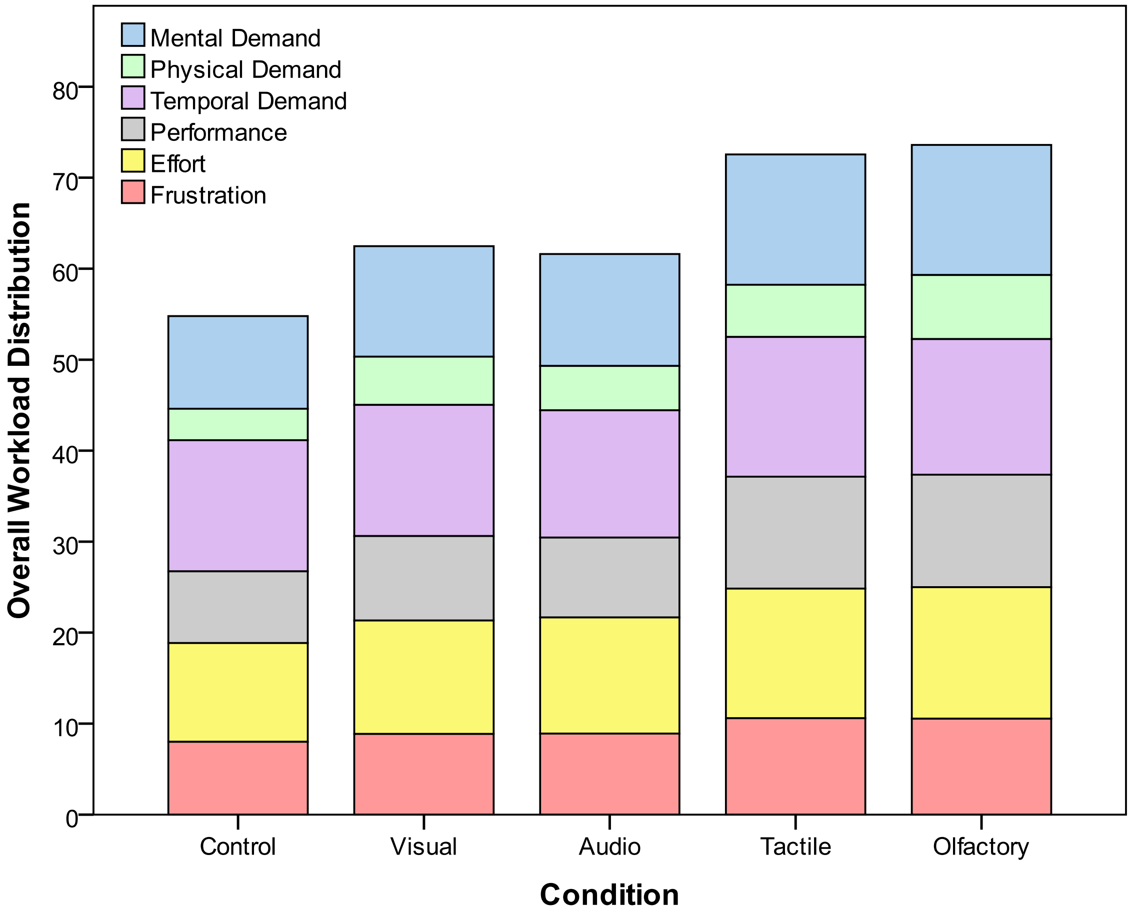


Figure 3.19: Composition of Overall Workload. This graph shows how the mean of each of the six 21-point NASA-TLX scales contributed to the overall workload.

than males when responding to olfactory notifications; they responded more quickly, but there was not a difference in response accuracy. This might suggest that olfactory notifications are more salient to females, however further testing would be needed to fully understand this result.

3.6.7 Hypothesis 3.1 (Subjective Workload)

Hypothesis 3.1 was defined in Section 3.3.3 as “the sensory channel to which notifications are delivered will affect the subjective workload involved with the task”. As discussed in Section 3.2.5, Workload (WL) was measured by NASA-TLX. The overall workload is represented by the total of 6 workload component measures: Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Overall Performance (OP), Effort (EF) and Frustration (FR). The overall workload and its composition is shown in Figure 3.19.

Overall Workload

WL was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on WL ($\chi^2(4) = 41.74,\ p < .001$). *Post hoc* tests, which are provided in Table A.4 in Appendix A, identified two groups; one containing the control, visual and audio groups, and the other containing the tactile and olfactory conditions.

Mental Demand

MD was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on MD ($\chi^2(4) = 34.60,\ p < .001$). Table A.5 shows the *post hoc* pairwise comparisons which revealed two groups; one containing the control condition and the other the audio, tactile and olfactory conditions. The visual condition was significantly different to the visual and olfactory conditions, but not from the control or audio conditions.

Physical Demand

PD was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on PD ($\chi^2(4) = 41.74,\ p < .001$). *Post hoc* pairwise comparisons are shown in Table A.6 which revealed that the tactile and olfactory conditions were significantly different from the control, while the visual and audio conditions were not significantly different from either group.

Temporal Demand

TD was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on TD ($\chi^2(4) = 9.56,\ p < .05$). However, Table A.7 showed that there were no significant differences between the conditions.

Performance

OP was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on OP ($\chi^2(4) = 38.30,\ p < .001$). *Post hoc* pairwise comparisons are shown in Table A.8 which identified two groups; one containing control and audio, the other tactile and olfactory. The visual condition was not significantly different from any of the other conditions.

Effort

EF was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on EF ($\chi^2(4) = 43.10,\ p < .001$). Table A.9 again shows two groups, one containing the control and visual conditions and the other containing the tactile and olfactory conditions. The audio condition was only significantly different from the olfactory condition.

Frustration

FR was evaluated using the non-parametric repeated-measures Friedman's ANOVA, which found that the condition had a main effect on FR ($\chi^2(4) = 15.76,\ p < .01$). *Post hoc* pairwise comparisons were carried out to determine the exact nature of the effects. Table A.10 shows that significant differences only existed between the control condition and the tactile and olfactory conditions. The visual and audio conditions were not significantly different from either group.

Assessment

The assessment of the subjective workload shows that a higher subjective workload was associated with the tactile and olfactory conditions compared to the visual and audio conditions. In addition, the visual and audio conditions were generally considered to produce a comparable workload to the control, suggesting that additional workload introduced by visual and audio notifications was negligible. These results support hypothesis 3.1, and will be explored further in the discussion in Section 3.7. Additional graphs and tables for the NASA-TLX data are provided in Appendix A.3.

3.6.8 Hypothesis 3.2 (Perceived *vs.* Actual Performance)

Hypothesis 3.2 was defined in Section 3.3.3 as "the perceived performance will not correlate to actual performance". Performance must be analysed for both the primary

Table 3.8: Table showing the correlations between perceived performance and actual performance in the primary and secondary tasks. As the NASA-TLX performance measure uses lower values for better performance, it was inverted for the analysis.

Condition	Primary Task		Secondary Task	
	Spearman's ρ	p	Spearman's ρ	p
Control	.64	<.001	—	—
Visual	.68	<.001	−.16	.44
Audio	.77	<.001	.49	<.05
Tactile	.49	<.01	.26	.20
Olfactory	.31	.126	−.17	.41

and secondary tasks, as the NASA-TLX performance measure is unable to distinguish between the two activities.[16]

The performance measure used for the primary task was the number of cards matched, and for the secondary task the percentage of correct responses was used. As shown in Table 3.8, participants appeared to assess their performance based on the number of cards they matched and not on the secondary task. Therefore the participants seemed to be quite effective at measuring their performance, as the correlations were both strong and significant for all of the conditions excluding olfactory.

The evidence does not support the hypothesis that participants will not be effective at assessing their performance. Instead, the evidence suggests that self-assessments of performance were quite accurate.

3.6.9 Hypothesis 4.1 (Visual LP & Reaction Time)

Hypothesis 4.1 was defined in Section 3.3.4 as "a higher visual learning preference will correlate to improved reaction times to visual notifications". Table 3.9 shows that no correlation was found; where the results were significant, the correlation value (Spearman's ρ) was small. The evidence instead suggests there is no relationship between visual learning preference and the speed of response to visual notifications.

[16] As discussed in Section 3.5, participants were asked to consider their overall performance, not performance on one task or the other.

Table 3.9: Modal learning prefence correlations to visual notification properties.

MLP	Reaction Time	% Correct	% Incorrect	% Missed
Visual	0.16	0.341*	−0.425*	0.049
Audio	−0.233	0.392*	−0.186	−0.315
Kinaesthetic	−0.145	0.038	−0.161	0.159

Note: Table shows Spearman's Rho (ρ). [*] = Significant to $p < .05$ (one-tailed).

Table 3.10: Modal learning prefence correlations to aural notification properties.

MLP	Reaction Time	% Correct	% Incorrect	% Missed
Visual	0.174	−0.222	0.236	0.202
Audio	−0.352*	0.221	−0.154	−0.292
Kinaesthetic	−0.187	−0.026	−0.13	0.074

Note: Table shows Spearman's Rho (ρ). [*] = Significant to $p < .05$ (one-tailed).

Table 3.11: Modal learning prefence correlations to tactile notification properties.

MLP	Reaction Time	% Correct	% Incorrect	% Missed
Visual	0.285	0.024	−0.213	0.382*
Audio	−0.198	0.003	0.099	0.189
Kinaesthetic	−0.099	0.066	−0.119	0.263

Note: Table shows Spearman's Rho (ρ). [*] = Significant to $p < .05$ (one-tailed).

Table 3.12: Modal learning prefence correlations to olfactory notification properties.

MLP	Reaction Time	% Correct	% Incorrect	% Missed
Visual	−0.063	0.127	−0.279	0.024
Audio	−0.347*	0.013	0.027	−0.032
Kinaesthetic	0.026	0.077	0.049	−0.078

Note: Table shows Spearman's Rho (ρ). [*] = Significant to $p < .05$ (one-tailed).

3.6.10 Hypothesis 4.2 (Visual LP & Response Accuracy)

Hypothesis 4.2 was defined in Section 3.3.4 as "a higher visual learning preference will correlate to improved response accuracy for visual notifications". Table 3.9 shows that there was a significant correlation between the number of correct ($\rho = .34\ p < .05$) and incorrect ($\rho = -.43\ p < .05$) notification responses. While this is the trend that was anticipated by the hypothesis, the Spearman correlation values for both results are not particularly high. This result will be discussed in Section 3.7.

3.6.11 Hypothesis 4.3 (Aural LP & Response Time)

Hypothesis 4.3 was defined in Section 3.3.4 as "a higher audio learning preference will correlate to improved reaction times to audio notifications". As shown in Table 3.10, there was a significant correlation between audio learning preference and reaction time ($\rho = -.35,\ r < .05$). However, much like Hypothesis 4.2 (Section 3.6.10) the Spearman correlation value is quite low, suggesting the correlation is quite weak. This result will be discussed in Section 3.7.

3.6.12 Hypothesis 4.4 (Aural LP & Response Accuracy)

Hypothesis 4.4 was defined in Section 3.3.4 as "a higher audio learning preference will correlate to improved response accuracy for audio notifications". Table 3.10 shows that there were no significant correlations between the accuracy of responses to aural notifications and audio learning preference. In this case, the evidence does not support the hypothesis.

3.6.13 Hypothesis 4.5 (MLP & Response Time)

Hypothesis 4.5 was defined in Section 3.3.4 as "tactile and olfactory notification reaction times will not correlate to any of the MLP measures". However, Table 3.12 shows that there was in fact a weak correlation between the audio learning preference and olfactory reaction time ($\rho = -.35,\ p < .05$). No other correlations were found, and this correlation itself is quite weak; therefore this result is somewhat inconclusive, and is discussed further in Section 3.7.

3.6.14 Hypothesis 4.6 (MLP & Response Accuracy)

Hypothesis 4.6 was defined in Section 3.3.4 as "tactile and olfactory response accuracy will not correlate to any of the MLP measures". Table 3.11 shows that a significant correlation did exist between visual learning preference and the number of notifications missed ($\rho = .38,\ p < .05$). No other correlations were found. The results of this test are somewhat inconclusive and are discussed in more detail in the following section.

3.7 Discussion

The results of the study identified many of the differences between the sensory channels, but also revealed that they are surprisingly similar in many ways. This section will explore the results and their implications. Each of the research questions is discussed separately: disruption in Section 3.7.1, effectiveness in Section 3.7.2, subjective perceptions in Section 3.7.3 and MLP in Section 3.7.4. Also included is a reflection on the validity of the experiment in Section 3.7.5.

3.7.1 Disruption

The basic performance measure, average matches per game, showed that performance dropped significantly in the experimental conditions, as expected. However, it also showed that the visual and audio notifications allowed the highest performance, and olfactory the worst. At a high level this seems to suggest that the tactile and olfactory notifications were more disruptive than the visual and auditory ones. It may also be possible that the novelty value of tactile and olfactory notifications caused some reduction in activity rate that resulted in poorer performance; if so, this would not demonstrate that these modalities were actually more disruptive. Hypothesis 1.2 and 1.3 provided additional insight into the behaviour of the player in these conditions.

Activity rate appeared to slightly (but not significantly) increase in the visual and audio conditions, which can be seen in Figure 3.15. None of the conditions were significantly different from the control, but there were significant differences between the olfactory condition and the visual and audio ones. A difference was also found between the tactile and audio conditions. This is surprising: the activity rate does not change significantly from the control, yet the performance dropped significantly compared to the control. This suggests that a change in activity rate is not related to the lower performance, although the lower activity rate in the olfactory and tactile conditions may be responsible for the lower performance compared to the visual condition. In other words, the lower activity rate does not explain the drop in performance across all the conditions; however it might explain the drop in performance between the conditions.

Superfluous views per turn (SVpT) was the metric devised to provide a greater insight into the disruption caused by notifications, representing the error rate in the primary task. This measure seems to more accurately reflect the drop in performance, as there was a significant increase in SVpT in all the experimental conditions compared to the control condition but no significant differences between experimental conditions. The SVpT measurements seem to suggest that there was no real difference in the error rate across the experimental conditions. In other words, if disruption is measured by the increase in error rate, then all the modalities were similarly disruptive.

The model of interruption discussed in Section 2.4.1 notes that there is a process of stopping a task and building a mental snapshot, which is then used to resume the task. The results appear to suggest that the sensory channel used to receive a notification does not have an effect on this process, and therefore does not affect disruption. The additional drop in performance associated with olfactory and tactile notifications may be attributed to longer breaks in activity; the slightly lower activity rates shown in Figure 3.15 and the longer response times in Figure 3.17 seem to support this hypothesis, but the measurements are too coarse to confirm it. A further study would be needed to identify if the additional performance drop should be attributed to increased disruption or decreased activity.

3.7.2 Effectiveness

The different sensory channels used demonstrated very different performance properties. The most straightforward measurement was a measurement of response speed; the time between notification delivery starting and the notification being responded to. Therefore, this measurement includes the time taken to deliver; this is important is it reflects the real-world reaction time to such notifications.

The results for reaction time showed that all the sensory channels were significantly different. Visual notifications were the quickest, followed by Audio which was only slightly slower. For both, most notifications were resolved in around 1.5 seconds. Tactile notifications were much slower than audio notifications, and as they were both based on sound and delivered notifications of comparable length, this was a surprising result. This suggests that there may be some sort of delay introduced by the tactile modality, *e.g.* participants might take longer than anticipated to interpret them. Unsurprisingly, the olfactory notifications took much longer to deliver than the other methods. It is hypothesised that this is due to the nature of smell interaction, and not due to additional time spent processing the notifications; unfortunately, this cannot be verified with the current measurements.

All of the sensory channels were effective at receiving notifications; the poorest performer was olfactory, which was correctly responded to 71% ($\sigma = 16$) of the time. This is much higher than anticipated, and given that olfactory interaction has unique properties and application areas (*e.g.* it is the only modality tested that can be used to interact with a deaf-blind person without physical contact) it should not be discounted. Tactile interaction demonstrated a similar performance of 75% ($\sigma = 19$), and the pairwise comparisons shown in Table A.1b revealed that the two modalities were not significantly different in terms of correct responses.

However, there are interesting differences between olfactory and tactile when examining the number of notifications incorrectly acknowledged and not acknowledged. As shown in Figure 3.18, olfactory notifications were much more likely to go unacknowledged than tactile notifications, while tactile notifications were much more likely to be incorrectly responded to. This suggests that tactons are highly salient but difficult to interpret, and that olfactory notifications are more subtle but are easier to understand.

As expected, visual and audio notifications were much more likely to be correctly acknowledged. However, one surprising result was the high variance in the audio condition (*e.g.* correct responses in the audio condition were $M = .88$, $\sigma = .24$, while in the visual condition $M = .96$, $\sigma = .06$). This can most likely be attributed to the different modalities that made up the audio condition; it is clear that one was less effective than the others. Exit interviews strongly suggested that the source of the variance was the Earcon condition. The validity of this aspect of the experiment is discussed further in Section 3.7.5.

The final hypothesis considered the role of gender in the olfactory condition. While no effect was found on the response accuracy, a significant effect was found on the response speed. This is particularly interesting because the response time is diluted by the delivery time of the olfactory notifications, which is expected to be roughly constant. It would be interesting to know why females outperformed males in this case; did they notice the notifications quicker (which would mean that they were more salient for females), or were they simply faster at interpreting them? Unfortunately, this question cannot be answered with the current measurements; a way to split up the delivery time from the interpretation time would be required to fully understand this phenomenon. This result suggests it may also be interesting to look for the effects of gender on the other notification modalities.

The results generally show that visual and audio methods were much more effective than the tactile and olfactory ones. However, the performance of the olfactory and tactile methods was not so poor that they should never be used as notification modalities. Their unique properties, *e.g.* the privacy that tactile can offer, suggest unique application areas. While most interactions should use the visual and audio modalities, other methods such as tactile and olfactory should not be discounted.

3.7.3 Subjective Measures

The subjective measurement was based around the NASA-TLX subjective workload assessment. The results presented in Section 3.6.7 (and additionally in Appendix A.3) showed that there were generally two groups; one containing the control, visual and audio conditions and the other containing the tactile and olfactory conditions. The olfactory and tactile group almost universally exhibited higher workload ratings, with the sole exception of temporal demand, which was shown to be the same across all conditions. This seems to verify the suggestion that participants found olfactory and tactile notifications harder to respond to than the others.

It is also interesting to note that the addition of the visual and audio secondary tasks did not result in a significantly higher workload. Speier *et al.* [189] suggested that some tasks were so simple that few cognitive resources were allocated to them, resulting in lower performance in those tasks. The addition of interruptions would increase the mental demand of the task and in turn the resources allocated, resulting in higher performance. This does not occur here; as shown in Section 3.6.1, performance in the visual and audio tasks decreased compared to the control. It does appear that subjectively at least, participants felt the additional demand of the visual and audio secondary tasks was negligible.

Table 3.8 suggested that subjective workload was mainly based on the primary task, shown by the correlations between subjective performance and performance in the primary task. Subjective performance and actual performance did not correlate for either of the performance measures in the olfactory condition, which suggests that some participants were unable to accurately assess their performance. The most likely explanation for this is that participants were less confident in the olfactory condition, possibly due to the novelty of the medium.

3.7.4 Modal Learning Preference

Research question 4 aimed to analyse any potential relationships between MLP and real-world modal performance. If a reliable relationship between performance and MLP could be discovered, it may be possible to predict notification performance based on a simple assessment (*i.e.* a VAK or VARK survey). The most promising application of such a discovery would be in pre-configuring a multimodal system to favour the 'preferred' modalities.[17]

There has been little work attempting to link MLP to a more general modal performance measure. In fact, linking modal learning preference to a particular style of education has been highly controversial. Sharp *et al.* claimed "VAK, together with many of the ideas surrounding and underpinning it, is an educational minefield" [184, p. 95]. To be convincing, any findings from this study would need to be strong and reliable; unfortunately, the majority of the statistical tests were inconclusive.

Section 3.6.10 showed that the number of correct responses in the visual condition correlated significantly with visual learning preference, although the correlation coefficient was not particularly high. A similar result was found in Section 3.6.11 for the reaction time to audio notifications. These correlations are significant and appear to travel in the expected direction; however, similar correlations were found in Sections 3.6.13 and 3.6.14 which are not in line with predictions. These results, along with the number of measures being examined, suggest that there is a high chance of committing a type I error.

Another factor may play a role, however; as noted in Section 3.7.2, there was high variance in the audio condition which may be due to the way three different modalities were used to create a single 'audio sensory channel' condition (the same was done for visual, but the results show little variance in the visual data). Would the results be different if the modalities were instead examined separately?

[17] I say 'preferred' because it is unclear if MLP is conscious, or if there is a correlation between MLP and the preferred modality for notification delivery.

Table 3.13: *Post hoc* examination of correlations between modal learning preference and individual modality performance measurements.

		Reaction	Percentage of Notifications		
		Time	Correct	Missed	Incorrect
Text	Visual LP	.111	.388	−.189	−.421
	Audio LP	−.462	.284	−.409	.281
Pictograms	Visual LP	.621*	−.142	.567	−.283
	Audio LP	.203	.650*	−.279	−.557
Abstract Visual	Visual LP	−.034	.542	—	−.542
	Audio LP	−.322	.401	—	−.401
Speech	Visual LP	.303	.052	−.138	.069
	Audio LP	.042	−.157	−.069	.276
Auditory Icons	Visual LP	.068	−.280	.280	.280
	Audio LP	−.707*	.280	−.280	−.280
Earcons	Visual LP	.287	−.362	.451	.212
	Audio LP	−.703*	.736*	−.522	−.604*

Note: LP = Learning Preference. Table shows Spearman's Rho (ρ), with $n = 9$ for each entry. [*] = Significant to $p < .05$ (one-tailed). No entry exists for missed abstract visual notifications because all 9 participants responded to all the notifications in this condition.

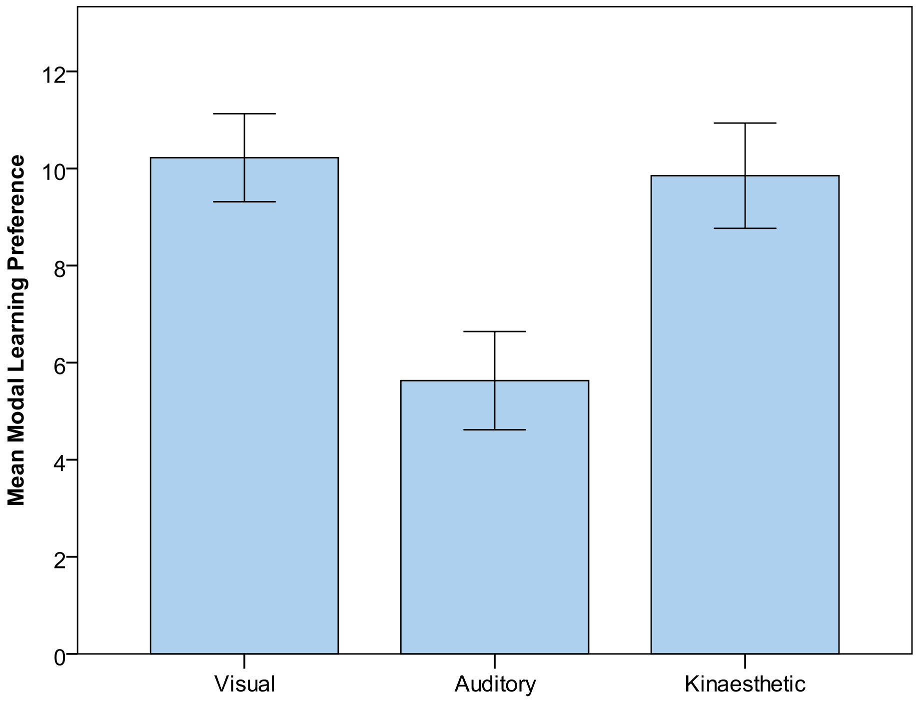


Figure 3.20: Graph showing the mean modal learning preference ratings for the experimental participants. Modal learning preference ranges from 0-20 and error bars show 95% confidence intervals.

This was examined *post hoc* and is shown in Table 3.13. In this table the number of participants for each entry is only 9, compared to 27 for the full analysis. It is likely that the lower number of participants is why the visual modalities do not reach significance. The audio modalities tell a different story, with two strong correlations showing that response time improved with higher audio learning preference in the auditory icon ($\rho = -.71,\ p < .05$) and earcon ($\rho = -.70,\ p < .05$) conditions. Earcons, which are completely abstract, also showed improvement with higher learning preference ($\rho = .74,\ p < .05$). The earcon condition is completely abstract, so if a higher audio learning preference does confer an improved audio processing ability, it would make sense for this to manifest in the earcon condition. Another interesting note is that the effectiveness of pictographic notifications appeared to be linked more to audio learning preference than visual learning preference; reaction time increased with visual learning preference ($\rho = .62,\ p < .05$) and correct responses increased with audio learning preference ($\rho = .65,\ p < .05$).

Unfortunately, given the small number of participants and the likelihood of making a type I error, these statistics cannot be considered reliable. Another issue with the reliability of the data are that modal learning preference was not controlled during the experiment. As shown by Figure 3.20, there were few participants with high audio learning preferences, although it would appear that the visual and kinaesthetic participants represent a normal sample of the population (as the average for both is near to 50%). This leads to a knock-on effect as without controls, the participants were allocated randomly to conditions. With only 9 participants per modality for the statistics shown in Table 3.13 uneven groups could have a large effect on the results.

3.7.5 Threats to Validity

The primary threat to validity in the experiment lies in the choice of examining sensory channel and not individual modalities. As discussed in Section 3.2, the intention was to provide an overview of sensory channel performance; the different modalities were intended to serve as a representative subset of the interactions available in each sensory channel. However, the results of the experiment clearly showed that for the audio condition the differences between modalities were much larger than anticipated; conversely, there was little variance between the visual modalities.

It is clear from the results of the experiment that future studies based around this template should focus on the individual modalities and not the sensory channel they use to interact. In particular, earcons appeared to perform much more poorly compared to other audio interaction methods and account for the majority of the incorrect notification responses shown in Figure 3.18. In a similar fashion, the modal learning preference results were very different when the data were examined from a modality perspective instead of a sensory channel perspective. Despite this, the study successfully demonstrated the differences between visual, audio, tactile and olfactory notification methods, and provided useful information about the performance of the different sensory channels.

There were also some minor issues with the experimental design. The earcons used in the experiment varied only in one direction (timbre), while the tactons varied in two (rythm and roughness). This may have contributed to the poor performance of earcons compared to other audio modalities, although the earcons were the same as those of MgGee-Lennon *et al.* [146].

The three colours used in the abstract-visual condition may have caused an issue. When projected onto the off-white wall, the yellow and green notifications looking overly similar, which caused some confusion at the training stage of the experiment. However, the abstract notifications performed very well under experimental conditions so there does not appear to have been a knock-on effect from this.

The participant selection presents another issue; as this technology is intended for older people, the small number of older participants used in this experiment suggests that the results are not representative of the standard home care technology user. While not all home care technology users are older people, future work addressed this by including more older users.

3.8 Guidelines

The findings of this study will be of use to the developers of almost all types of multimodal technology, although the aim of the study was specifically to support multimodal home care systems. In this section the results of the study are refined into straightforward guidelines that can be applied in such settings.

- Assuming no significant impairment or background interference, visual and audio interaction methods provide the quickest and most effective ways to deliver information.

- Error rate is not affected by the sensory channel used by an interruption As the objective negative effects of an interruption appear not to be associated with the sensory channel it uses, there is no obvious reason as to why the 8 modalities demonstrated in this study cannot be deployed in the home.

- There is an additional subjective workload when using tactile and olfactory methods to deliver notifications. While the effect is small, visual and auditory notifications should be preferred in most circumstances.

- Tactile notifications may be difficult to understand, but are more salient than olfactory notifications.

- Olfactory notifications are not highly salient and are likely to be missed. When noticed however, participants had little trouble understanding their meaning.

- Olfactory notifications take longer to deliver, and should not be used for important notifications if other modalities can be used.

- Olfactory notifications appear to be more effective with female participants; however, the performance difference is not large enough that olfactory notifications should never be used to deliver information to males.

These are the primary guidelines resulting from the study presented in this chapter. However, the experiment also provided raw performance data on each of the delivery methods which are of value when attempting to design multimodal notifications. There were also several partial findings, such as the observed ‘tactile lag’, which will require further study to be fully understood.

No guidelines have been presented based on age or modal learning preference. There were too few participants to allow for a reliable analysis on any effects related to the participant’s age. As discussed in Section 3.7.4, the results for the modal learning preference parts of this study were promising, but not convincing enough to use as the basis for dependable guidelines.

3.9 Conclusions

Thesis Question 1 considered which methods of interaction would be most appropriate for home care technology. The work in this chapter showed that the answer to this question is not straightforward. While visual and auditory notifications are the most effective, alternative methods such as tactile and olfactory are not ineffective, although they do have very different properties. As long as these properties are well understood, these modalities could be used to help overcome some of the challenges faced by the designers of multimodal ALTs. For example, Vastenburg [202] argued that low-importance notifications should be delivered using low-salience methods, and that users did not care if they did not receive them. Olfactory notifications are likely to be ideal for this scenario, but are unlikely to be appropriate for delivering urgent and important notifications as they are slow and likely to be missed. In general, the results of the study show no clear reason why any of the sensory channels should be excluded from home care technology; instead the results suggest that multiple sensory channels should be used as they can be used to solve a wide range of problems that might be found in a home care setting.

These results are not sufficient to answer Thesis Question 2, and additional work was needed to form a better understanding of how these modalities affect users and ongoing tasks. Section 3.7.2 demonstrated the value of being able to differentiate between the *delivery time* of a notification and the *processing time* of a notification. That information could be used to clarify the nature and source of the disruptive effects discussed in Section 3.7.1. Another interesting area of study is the potential disruptive effects of *distractions*. A distraction is likely to occur in situations where the user does not want to receive the reminder, or if the reminder is delivered to a cohabitant or houseguest. It is well-known that the home tends to be a multi-user environment [145, 207], so non-users of reminder technology are likely to find the notifications distracting. Sanders & Baron [180] noted that distractions caused similar negative effects to normal interruptions, and distractions are known to increase the chance of a fall at home [21]. Addressing Thesis Question 2 required a study that: *(1)* focussed on each modality individually, *(2)* included older people in the study, *(3)* considered the effects of distractions and *(4)* provided fine-grained data on the process of receiving a notifications in a given modality. That study is presented in Chapter 4.

In conclusion, this chapter has addressed Thesis Question 1 by carrying out a baseline study of notifications delivered to different sensory channels. The results have shown that the range of modalities that can be used in the home is much larger than expected, although different modalities should be used in different situations. The guidelines provided in Section 3.8 show that there is a clear need for additional data to provide a comprehensive understanding of how different modalities can be used to support care at home. This is addressed in Chapter 4, which presents a study designed to answer Thesis Question 2 that considers the effects of notifications delivered in different modalities in greater detail.

Chapter 4

Further Study of Notification Modality

Thesis Question 2 was *"how do different forms of notification delivery affect users?"*. Chapter 3 presented a study into the effects of delivering notifications using the visual, audio, tactile and olfactory sensory channels. This provided useful baseline information regrading the properties and applications of different types of notification delivery methods. However, this information was not sufficient to answer Thesis Question 2 and Chapter 3 concluded that additional research was needed that considered individual modalities (instead of grouping them by sensory channel) and examined the effects of distractions in different modalities. It was also noted that to fully answer Thesis Questions 1 and 2, studies should use a mixture of younger and older participants to provide a more reliable understanding of how different modalities affect users and ongoing activities.

This chapter presents a study that addressed Thesis Question 2 by expanding on the experiment carried out in the previous chapter. The aims of this study are laid out in Section 4.1. Section 4.2 details the design of the experiment, followed by Section 4.3 which refines the research questions into a set of testable hypotheses. Section 4.4 describes the participants who took part in the study and Section 4.5 describes the procedure that was followed when the study was carried out. The results of the study are presented in Section 4.6, followed by a discussion in Section 4.7. The primary findings of the study are refined into useful guidelines in Section 4.8 and this chapter closes with the conclusion in Section 4.9.

4.1 Aims

Chapter 3 examined several modalities and provided useful baseline data about the disruptiveness and effectiveness of notifications using those channels. However, Section 3.7 concluded that the modalities should be examined separately; while the three visual modalities (text, pictograms and abstract-visual messages) appeared to have similar properties, the results suggested differences between the 3 audio modalities (speech, auditory icons and earcons). One of the aims of this study was to examine the modalities separately to form a better understanding of how they could be used in the home.

The study also addressed some important points raised in Section 3.7, one of which was that the effects of age on the results should be examined. Existing work has shown that naturally-occurring sensory and cognitive decline with ageing [167, 114, 129, 169, 188], so older participants are not expected to perform as well as younger participants in the tasks. Older users may also be more disrupted by interruptions, but it is not currently known if modality and age will interact in any way.

Another point raised was the effect of distractor notifications, *i.e.* notifications which have no value to the receiver. Existing work has not only shown that distractions can produce the same negative effects as interruptions [180], but are also frequently cited as a cause of falls in the older population [21]. There are many ways for a reminder system to create distractions. It might deliver a notification during an important and cognitively demanding activity. As the home is often a multi-user environment [144], notifications intended for one resident might simply be unwanted distractions to co-habitants. In some cases, delivering distractor notifications may be unavoidable; as such the effects they will have must be well-understood if home reminder technology aims to balance acceptability and effectiveness.

The main aims of the study were to address the following research questions:

Research Question 1

What are the effects of modality and age on the disruptiveness of a notification to an ongoing task?

Research Question 2

What are the effects of modality and age on notification effectiveness?

Research Question 3

How does age and modality affect the subjective workload associated with a notification?

Research Question 4

In what ways do distracting notifications (*i.e.* notifications that serve no purpose) affect disruption and effectiveness? Are some modalities less distracting than others?

Together, these research questions address Thesis Question 2. This is an important step that must be completed to allow Thesis Question 3 to be answered. Considering the effects of age will also provide an additional insight into which modalities are appropriate for the home, adding to the findings of Chapter 3 and improving the answer to Thesis Question 1. The following section discusses how the study was designed to answer the four research questions.

4.2 Design

The design used in the first study (see Section 3.2) was re-used in this study. This design involved asking participants to play a simple card-matching game. Participants would be interrupted with a secondary task during the game, with the interruption using a different sensory channel in different conditions. The effects of the sensory channel on performance in the primary and secondary tasks was measured to provide an understanding of how different sensory channels affected the disruptiveness and effectiveness of notifications. Subjective data were also gathered to provide an insight into perceptions of the different sensory channels.

Several modifications were made to this design to address the research questions outlined in Section 4.1. The primary task used in the game is discussed in Section 4.2.1, the secondary task in Section 4.2.2, and the design of the notifications in Section 4.2.3.

The study was a mixed-models design, with age as a between-groups variable and modality and notification relevance as within-groups variables; these are discussed in Section 4.2.4. Compared to the first study, several changes were made to the way that data were gathered to provide an insight into the process of stopping the card-matching game, dealing with a notification and then resuming the game; these changes are described in Section 4.2.5, which specifies the dependent variables measured in the study. Section 4.2.6 outlines potential confounding variables that might interfere with the experiment and how they were controlled or monitored during the study.

4.2.1 Primary Task

Section 3.2.1 justified the selection of a simple card-matching game (generally known as Concentration) as the primary task for that study. This study follows a design that is very similar to that used in the first study, in which the card-matching task performed well and provided a range of useful measurements. Therefore, the primary task did not need to be modified.

Each game of Concentration included 24 cards arranged in a $4 * 6$ grid. The cards used drawings with simple verbal labels as shown in Figure 3.2. Games had a 60-second time limit, with remaining time shown at the bottom of the screen in each game. Due to the increased number of conditions in the experiment (each experiment required 9 conditions, while the previous experiment had 5) the number of games played in each condition was reduced to 4 (from 5 in the previous study).

4.2.2 Secondary Task

In the previous study, participants were asked to respond to three different notifications by pressing one of three buttons, each with a matching context where possible (*e.g.* if the word “Telephone” appeared on the screen, the button labelled “Telephone” should be pressed). The original secondary task could not be re-used for two reasons: *(1)* there was no straightforward way to introduce distractor notifications into the task, and *(2)* the original task required considerable training time, which had to be minimised to allow for the larger number of conditions.

The primary task was modified to make it simpler by replacing the three buttons with a single yellow button as shown in Figure 4.1. No context was provided on the button to prevent confusion (as noted in the following section, the abstract-visual condition used green, red and blue). As in the first study, three notifications were used in each modality corresponding to ‘lights’, ‘heating’ and ‘telephone’. One of the messages was selected at random to be the *target notification*, which participants were asked to acknowledge by pressing the yellow button. The other two notifications were called *distractor notifications*, which participants were asked to try and ignore entirely.

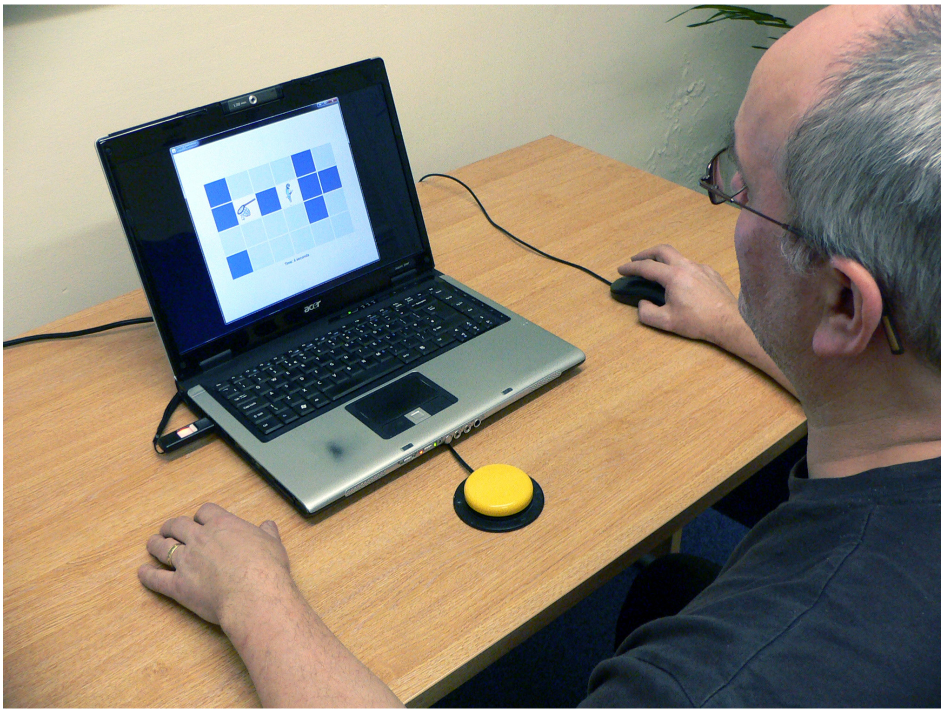


Figure 4.1: The experimental setup in the second study only used 1 button, which did not provide any context for the notifications unlike the buttons used in the first study.

In each condition 12 notifications were delivered (3 per game), of which 6 were target notifications and 6 were distractors. The notifications were randomised over all the games as described in Section 3.2.2, so it was possible for players to play a single game that included only target or distractor notifications. This helped to prevent players from predicting which notifications would be next. The delivery timing for notifications was handled in the same was as it was in the first study, which is explained in Section 3.2.2 and Figure 3.4.

4.2.3 Notification Design

The notifications used in this study were mostly the same as the ones used in the previous study: text, pictograms, abstract-visual messages, speech, auditory icons, earcons, tactons and olfactory. As described in Section 3.2.3, 3 messages were developed for each of the modalities, ideally with the context of 'heating', 'lights' and 'telephone'. Based on the results and feedback from the previous study some changes were made to the notifications.

For the abstract-visual condition, the colours used were changed to increase the contrast between them. The new colours used were red, green and blue. The colours were tested against the wall that would be used in the experiment to make sure that the off-white

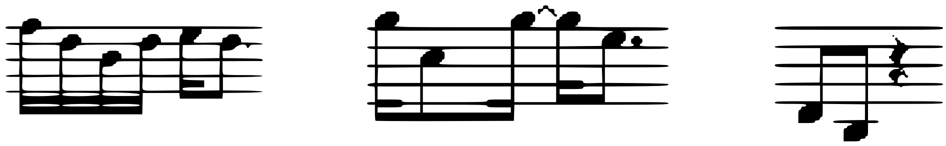


Figure 4.2: The earcons used in this study, from McGookin & Brewster [148, p. 139]. From left to right these were called high, medium and low intensity (HI, MI and LI).

paint would not affect perception of the colours. The yellow button shown in Figure 4.1 was chosen simply because it was not similar to the colours used in this condition.

The earcons used in the first study only varied in timbre (the sound of the instrument), while the tactons varied in roughness and rhythm. Section 3.7 noted that this may have meant the earcons were too similar to each other, which could have reduced their effectiveness during the study. To ensure that this was not the case, new earcons were selected from the work of McGookin & Brewster [148] which varied in rhythm and timbre. The earcons are shown in Figure 4.2 and were played using standard MIDI sounds with the same volume. The length of the sound files for the earcons was 1-2 seconds.

The other notifications and their delivery methods were the same as those used in the first study. The notifications used are summarised in Table 4.1, and more information can be found in Section 3.2.3 on why these modalities were selected for the study and a description of the hardware used.

4.2.4 Independent Variables

There were 3 independent variables in the study: age, modality and notification relevance. Age was a between-groups variable, while modality and notification relevance were repeated-measures variables.

Two age groups were used in the experiment, called the younger group and the older group. The younger group was limited to participants aged from 16-30, while the older user group was limited to participants over 50. Full details of the study's participants is provided in Section 4.4.

There were 8 modalities used in the study: text, pictograms, abstract-visual messages, speech, auditory icons, earcons, tactons and olfactory notifications. Each of these was configured as described in Section 4.2.3. Unlike the first study, participants were asked to complete a condition for all 8 modalities. There was also a control condition where no

Table 4.1: Configuration of the modalities used in the second study.

Modality	Message			Delivery Timing	Notes
	Heating	Lights	Telephone		
Text	"Heating"	"Lights"	"Telephone"	Continuous	Delivered at the top of the game area.
Pictogram	ISO-7000 [107] Thermometer	ISO-7000 [107] Light	ISO-7001 [106] Telephone	Continuous	Delivered at the top of the game area.
Abstract Visual	Red #FF0000	Lime #00FF00	Blue #0000FF	Continuous	Coloured light delivered into peripheral vision.
Speech	Spoken "Heating"	Spoken "Lights"	Spoken "Phone"	Finite	Delivered via audio, voice was synthetic 'Heather' from the Scottish Voice.
Auditory Icon	Gas Ignition	Light Switch Click (x2)	Dialling Beeps	Finite	Taken from a royalty-free sound-effects archive.
Earcon	HI Synth Brass (Gen. MIDI #062)	MI Bright Acoustic Piano (Gen. MIDI #001)	LI Acoustic Grand Piano (Gen. MIDI #000)	Finite	From McGookin & Brewster [148, p. 139], earcons varied in rhythm and timbre.
Tacton	multiLP	textLP	voiceLP	Finite	From Brown *et al.* [34], tactons varied in rhythm and roughness.
Olfactory	'Dark Chocolate'	'Riverside'	'Raspberry'	Delayed-continuous	Delivered using Vortex device. Smell names defined by the Dale Air catalogue.

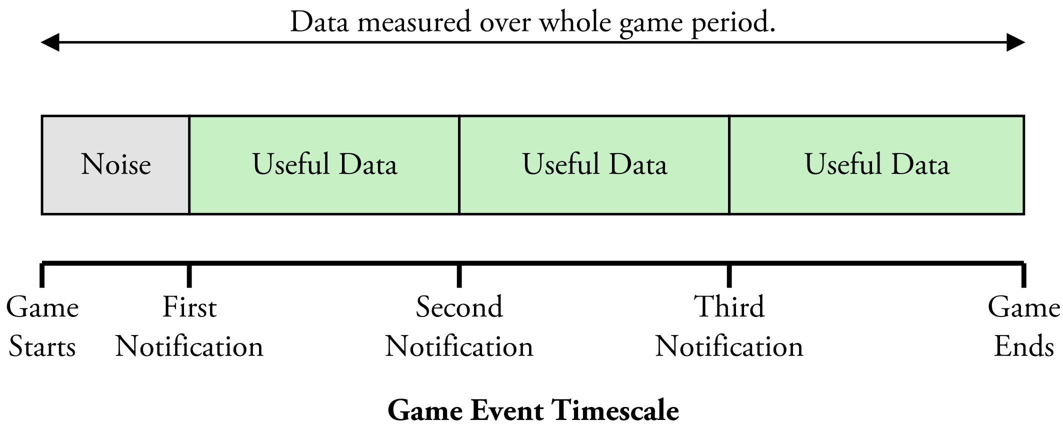


Figure 4.3: Diagram showing how data were gathered in the first study, which included data from the 'warming up' buffer at the start of each game.

modalities were delivered, for 9 conditions in total. These were delivered to participants in a random order. Additional details of the experimental procedure are provided in Section 4.5.

Notification relevance defines whether a notification would require a response. There were two levels for this variable: notifications would either require acknowledgement by pressing the yellow button (as described in Section 4.2.2) or would require no action at all. Notifications requiring an acknowledgement were called *target* notifications, while notifications to be ignored were called *distractor* notifications.

4.2.5 Dependent Variables

Section 3.2.5 outlined several dependent variables that were used in the first study to investigate the research questions, grouped under 'disruption', 'effectiveness' and 'subjective workload'. The three measurements for disruption (performance, activity rate and error rate) proved effective in the first study and were re-used in this study to measure both disruption. The NASA Task-Load Index (NASA-TLX) [79] form used in the first experiment was also useful for subjective workload data.

One of the shortcomings of the previous study was that the measurements of effectiveness were too coarse to provide an insight into the process of pausing and resuming the primary task. The root cause of this was the way that data were gathered; as Figure 4.3 shows, data were measured over the entire length of a game then averaged, including the long buffer zone at the start of each game.

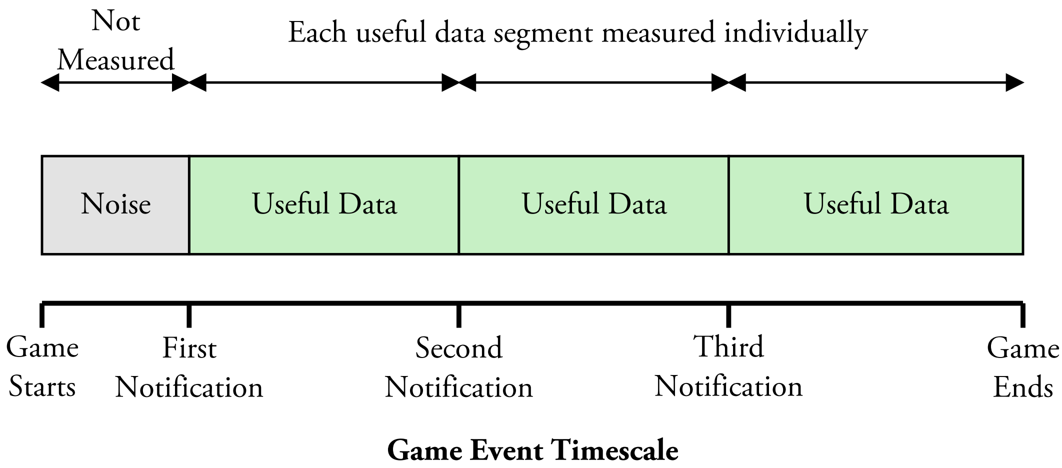


Figure 4.4: Diagram showing the difference in measurement methods adopted for Experiment 2, which eliminated a considerable amount of noise from the data.

The data of interest lies in the period immediately after each notification, which is diluted if measurements are taken over the whole game. To combat this each game was split into three segments, each of which started at the point of notification delivery and continued up to the start of the next notification as shown in Figure 4.4. Data from the warm-up buffer were discarded.

As well as providing cleaner data, examining the data in this way allowed the after-effects of target and distractor notifications to be analysed with the same design. However, this still left the problem of finding more fine-grained measurements that would provide a greater insight into the task-switching process. Section 3.7 noted that the ideal measurements would include a way to differentiate between the time taken to deliver a notification and the time taken to process and respond to the notification.

To measure the time spent processing and responding to notifications, a new measurement was developed called the *longest pause*. The longest pause is the longest break in activity in the time period following a notification, as shown in Figure 4.5. The longest pause could be measured for both the target and distractor notifications without asking the participant to take any kind of action; for distractors, participants were asked to ignore the notifications and continue as normal. A measurement of 'real' delivery time could then be calculated by taking the time difference between the start of notification delivery and the start of the longest pause. These measurements would provide the level of insight required to answer questions about the process of starting and stopping the primary task without requiring significant changes to the experimental design.

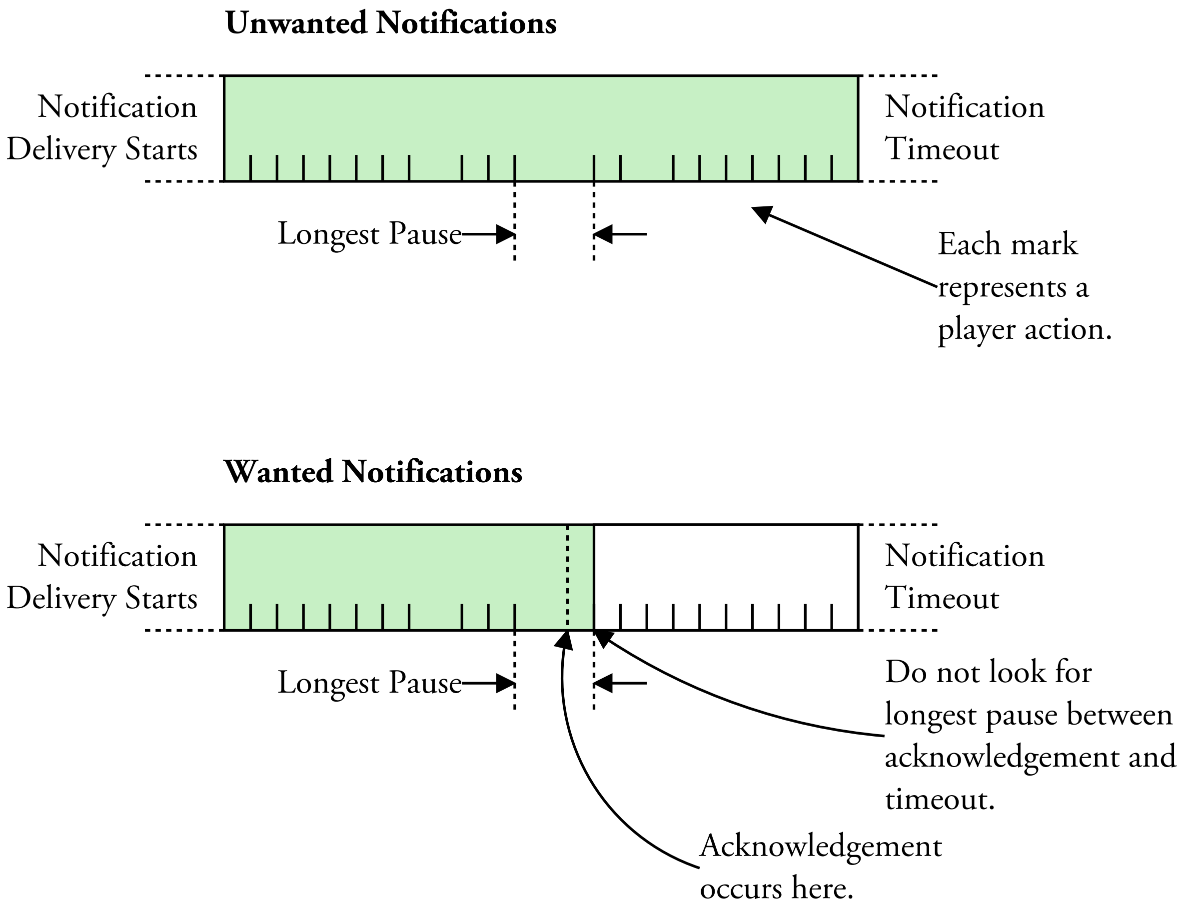


Figure 4.5: Diagram demonstrating how longest pause was measured. Each of the marks represents a player action, in this case a click. Over the measurement period (shown in green above), the time between clicks varied. Some clicks occurred quickly after the preceding click and others took longer. The *longest pause* is a measure of the largest time between clicks over the measurement period, which is taken to be the point at which the user stops to mentally and physically process the notification.

The other measure of effectiveness was response accuracy, which was simplified compared to the previous study. The previous study had 3 potential outcomes: correct response, incorrect response and no response. This study had 4 potential outcomes; correctly acknowledged, incorrectly acknowledged, correctly ignored and incorrectly ignored. The two correct responses were grouped and expressed as a percentage to provide a single overall assessment of response accuracy.

Summary of Dependent Variables

The majority of the measures are the same as the ones used in the first study, which were defined in Section 3.2.5. However, several modifications are described in this Section which were used to improve the fidelity of the measurements, and two additional measurements were created. To summarise, the full list of measurements that were taken in the study are as follows:

Disruption: Performance

Performance was measured using Cards Matched (CM), the same measurement used in the first study. This measurement ranged from 0 to 12 and represented real-world performance in the primary task.

Disruption: Activity Rate

Activity rate was measured as the number of clicks per second, similar to the measurement taken in the first experiment. However, the period of measurement was changed for this study to measure only the activity rate after each target/distractor notification.

Disruption: Error Rate

Superfluous Views per Turn (SVpT), defined in Section 3.2.5, was again used as a measurement of error rate. Like activity rate, this measurement was refined to only include data post-notification instead of over the whole game. The range for error rate is 0 to 2.[1]

Effectiveness: Response Accuracy

Response accuracy was measured by the number of target notifications correctly acknowledged and distractors correctly ignored, expressed as a percentage.

[1]In the original study, the upper limit had to include the first turn, which made the upper limit an asymptote defined by $\left[2 - \left(\frac{2}{t}\right)\right]$ where t is the number of turns in the game. However, in this study the first turn is unlikely to take place in any of the measurement periods, so the *worst-case* performance becomes 2.

Effectiveness: Processing Time

The processing time was measured using the *longest pause* measurement; the longest pause in activity post-notification, measured in seconds.

Effectiveness: Delivery Time

The deliver time was measured as the time in seconds between the start of notification delivery and the start of the longest pause.

Subjective Measures: Subjective Workload

The NASA-TLX form was re-used from the first experiment. This provides 7 subjective measurements relating to the tasks; overall workload, physical demand, mental demand, temporal demand, effort, perceived performance and frustration.

4.2.6 Confounding Variables

The previous study noted that the most important confounding variables were age, gender and background noise (see Section 3.2.6). In this study age was a between-groups independent variable and is addressed in Section 4.2.4. The previous study suggested that gender influenced the results of the smell condition, so an attempt should be made to balance the experimental participants by gender. The same precautions were taken to avoid background interference as they were in the first study, and the same demographic survey was administered at the start of the experiment (see Figure 3.13).

With older participants there was an increased risk of sensory impairment affecting the results of the study. The sensory self-assessment survey, show in Figure 3.13b, provided an insight into sensory impairments to help prevent significant impairments from impacting on the results. The participant information sheet (which is provided in Appendix B) specifically asked participants with significant sensory impairments not to participate in the study. As noted in Section 4.4, none of the participants (from both age groups) exhibited or declared a significant sensory impairment.

4.3 Hypotheses

In this section a set of hypotheses are outlined to explore the research questions provided in Section 4.1.

4.3.1 Research Question 1 – Disruption

Research question 1 asked "*what are the effects of modality and age on the disruptiveness of a notification to an ongoing task?*". This was tested by measuring three variables: performance, activity rate and error rate, as defined in Section 4.2.5.

Performance – H 1.1

As discussed in Section 4.2.5, performance was measured by the number of *cards matched* in a game. The results of the first study (see Section 3.6) revealed that modality had an effect on performance, with performance dropping in the tactile and olfactory conditions. This trend is expected here as well, although when viewing the modalities individually (instead of grouping by sensory channel, as was done in Chapter 3), it might reveal differences between audio and visual modalities. Age is expected to be a factor in performance in the game, as short-term memory is shown to decline with age [179]; however, there is no reason to believe that there will be an interaction between age and modality. This provides the following hypotheses:

H 1.1.1 The modality used to deliver a notification will affect the number of cards matched.

H 1.1.2 The age of the participant will affect the number of cards matched.

H 1.1.3 There will be no interaction between modality and age on the number of cards matched.

Activity Rate – H 1.2

As discussed in Section 4.2.5, activity rate was measured by the number of clicks per second. The previous study (see Section 3.6) revealed that activity rate changes were unusual, as activity rate appeared to increase slightly to compensate for the additional work of the secondary task. Due to the differences between the way activity rate is measured in this study (discussed in Section 4.2.5) the control will not be included here; only the experimental conditions are included.

It is anticipated that there will be differences between age groups on activity rate, with younger participants producing a higher activity rate than the older participant group. In Section 3.7.1, it was hypothesised that the drop in activity rate was due to more time spent 'processing' the notifications. While modality is expected to have an effect on activity rate, it is not expected to be particularly strong. As with performance, no interactions are expected between age and modality on activity rate. These hypotheses are expressed as follows:

H 1.2.1 The modality used to deliver a notification will affect the click rate.

H 1.2.2 The age of the participant will affect the click rate.

H 1.2.3 There will be no interaction between modality and age on the click rate.

Error Rate – H 1.3

As discussed in Section 4.2.5, error rate was measured by the number of *superfluous views per turn*. The results of the first study (see Section 3.6) suggested that the sensory channel used would not affect error rate, and as such a similar result is expected when considering modality. Due to the effects of age on memory [179], the error rate is expected to be higher for older participants, but no interaction between age and modality is anticipated. This is shown by the following hypotheses:

H 1.3.1 The modality used to deliver a notification will not affect the number of superfluous views per turn.

H 1.3.2 The age of the participant will affect the number of superfluous views per turn.

H 1.3.3 There will be no interaction between modality and age on the number of superfluous views per turn.

4.3.2 Research Question 2 – Effectiveness

Research question 2 asked "*what are the effects of modality and age on notification effectiveness?*". Effectiveness was measured in three ways: response accuracy, delivery time and processing time. These measures were defined in Section 4.2.5.

Response Accuracy – H 2.1

As noted in Section 4.2.5, notification response accuracy is much easier to measure here than it was in the previous study. This is because there are only two possible outcomes; correct response and incorrect response. One one of these outcomes needs to be assessed, and so response accuracy is measured by the percentage of correct responses.

The previous study demonstrated that notification modality had a significant effect on the response accuracy, and this effect is expected to be recreated here. In particular, earcons are expected to perform poorly compared to other audio methods, as suggested in Section 3.2. As participants were expected to have no significant impairments, no effect of age is expected, nor interactions with modality. Therefore, the following hypotheses were tested:

H 2.1.1 Response accuracy will be affected by the modality used to deliver the notifications.

H 2.1.2 Response accuracy will not be affected by age.

H 2.1.3 Age and modality will not interact to affect response accuracy.

Processing Time – H 2.2

The previous study revealed an interesting phenomenon when responding to tactile notifications that was called the ‘tactile lag’; a very short delay in responding to tactile notifications. As the delivery hardware operated at the same speed as audio hardware, it was theorised that this tactile lag shows an increase in processing time for that modality. Due to the basic measurement taken, it was impossible to verify if a similar effect occurred in the olfactory condition, and there was no evidence of this in the visual or audio conditions. The *longest pause* measurement (described in Section 4.2.5) was designed to isolate the time spent thinking about and responding to a given notification. Based on the previous results, it is expected that modality will have an effect on processing time. Age is also expected to have an effect on processing time; the older participant group is likely to take longer to respond to the notifications. However, there is no evidence to suggest an interaction between the two factors. The hypotheses are as follows:

H 2.2.1 Processing time will be affected by the modality used to deliver the notifications.

H 2.2.2 Processing time will be affected by age.

H 2.2.3 Age and modality will not interact to affect processing time.

Delivery Time – H 2.3

As discussed in Section 4.2.5, the time between the initiation of delivery and the start of the longest pause can be used to more accurately measure the time taken to deliver a notification to a participant. The previous study already showed that olfactory notifications took significantly longer to deliver than others, but it also showed that audio notifications were slightly slower than visual notifications. The delivery time measurement is expected to be affected by the modality of a notification, with a *post hoc* pairwise comparison expected to show that olfactory has the longest delivery time by a considerable margin. Age is not expected to have an effect on the delivery time, not is it expected to interact with modality. The hypotheses are as follows:

H 2.3.1 Delivery time will be affected by the modality used to deliver the notifications.

H 2.3.2 Delivery time will not be affected by age.

H 2.3.3 Age and modality will not interact to affect delivery time.

4.3.3 Research Question 3 – Workload

Research question 3 asked "*how does age and modality affect the subjective workload associated with a notification?*". As in Chapter 3, this section uses data gathered by the NASA-TLX subjective workload survey.

The work in Chapter 3 found several differences between the sensory apparatus under NASA-TLX, mainly separating the control, visual and audio conditions from the tactile and olfactory conditions. Therefore, it is expected that similar effects will be observed when considering modalities instead of sensory apparatus. It is also anticipated that due to natural cognitive decline with age, the older participants will show higher workload scores than the younger participants. These predictions are expressed as the following hypotheses:

H 3.1 For younger participants, subjective workload will be affected by the modality used to deliver notifications.

H 3.2 For older participants, subjective workload will be affected by the modality used to deliver notifications.

H 3.3 Older participants will report higher subjective workloads than younger participants.

4.3.4 Research Question 4 – Distraction

Research question 4 asked "*in what ways do distracting notifications (i.e. notifications that serve no purpose) affect disruption and effectiveness?*". Distracting notifications were present through the experiments, as described in Section 4.2, this study examined the effects immediately after the delivery of the notification (as opposed to the effect over the whole game). Therefore, only the fine-grained measurements were used to evaluate this research question: activity rate, error rate and processing time.

Activity Rate – H 4.1

As discussed in Section 4.3.1, activity rate is considered a property of disruption and was measured by the number of *clicks per second*. Sanders & Baron [180] suggested that distractions will cause activity rate to rise; however the disruptions in the first study did not appear to affect activity rate (see Section 3.6.2). As such no significant differences were expected in activity rate following distractor and target notifications. In addition, no interactions were expected. This was tested with the following hypotheses:

H 4.1.1 Notification relevance will not affect click rate.

H 4.1.2 Notification relevance will not interact with modality to affect click rate.

H 4.1.3 Notification relevance will not interact with age to affect click rate.

Error Rate – H 4.2

Like activity rate, error rate is also considered a property of disruption measured by the number of *superfluous views per turn*. If distractions create the same effects as interruptions then notification relevance will not affect error rate, nor produce any interaction effects with age or modality. This was tested with the following hypotheses:

H 4.2.1 Notification relevance will not affect the number of superfluous views per turn.

H 4.2.2 Notification relevance will not interact with modality to affect the number of superfluous views per turn.

H 4.2.3 Notification relevance will not interact with age to affect the number of superfluous views per turn.

Processing Time – H 4.3

Processing time is a part of effectiveness and was measured using the *longest pause* measurement defined in Section 4.2.5. As target notifications required a response (pressing a button) while distractors were to be ignored, it is to be expected that the processing time for target notifications would be much longer. In addition, the hypothesised longer processing times associated with tactile and olfactory notifications (see Section 4.3.2) would suggest an interaction between the the notification function and modality. However, there is no reason to anticipate any interaction between age and relevance. The following hypotheses were tested:

H 4.3.1 Distractor notifications will produce a shorter 'longest pause' than target notifications.

H 4.3.2 Notification relevance and modality will interact to affect longest pause.

H 4.3.3 Notification relevance will not interact with age to affect longest pause.

Delivery Time – H 4.4

Delivery time is a part of effectiveness and was measured as defined in Section 4.2.5. Delivery time is measured as the time between the initiation of notification delivery and the start of the longest pause (where the longest pause is given as the point where the mental process of interpreting and responding to the notification takes place). Regardless of whether the information is relevant, it is expected that it will take the same amount of time to deliver to the participant. Therefore, the final set of hypotheses are:

H 4.4.1 Notification relevance will not significantly affect delivery time.

H 4.4.2 Notification relevance and modality will not interact to affect delivery time.

H 4.4.3 Notification relevance will not interact with age to affect delivery time.

4.4 Participants

There were two participant groups in the experiment: the younger group aged 18-30 and the older group aged 50+. The younger participants were university students who replied to an request for participants sent out to a student mailing list. There were 20 participants in the younger group (8 female and 12 male).

Older participants responded to a call for participants sent out to a mailing list related to the study. There was some difficulty in finding older males to take part in the experiment. A total of 16 participants were used in the older group (10 female and 6 male).

As participants were asked for their age range instead of their specific age, the mean age of participants in the two groups is not known. This shortcoming is discussed in Section 4.7.

4.5 Procedure

At the beginning of the experiment participants were given an information sheet and consent form (provided in Appendix B). Participants were then asked to take a short demographic survey to collect gender and age information, followed by a sensory-self assessment as shown in Figure 3.13.

At the start of the experiment participants were asked to carry out some practice games to familiarise themselves with the task. During the game a textual message would appear at the top of the screen that either say "please press the button" or "please ignore this message". The participants would play 3 of these practice games before continuing on to the experiment itself and were given the opportunity to repeat the practice if they did not fully understand the process.

There were 8 experimental conditions (one for each of the modalities shown in Table 4.1 and a control condition, where no notifications were delivered. These were delivered in a random order to the participants. The control condition simply asked the participant to play 4 games, during which no notifications were delivered. This would help to provide useful baseline performance data for the card-matching task.

The experimental conditions started with a training stage. This started by explaining the modality to the participant and setting up any hardware required for the condition (*e.g.* headphones). Before continuing, participants were allowed to make minor adjustments for comfort (*e.g.* volume). When the participant was ready, all 3 messages were shown to them in order. One of the 3 messages would be selected at random to be target notification. A screen would then reveal the target notification to the participant and instruct them to only acknowledge that notification and to ignore all others. Figure 4.6 shows that this screen included a button that would allow participants to repeatedly review the target notification until they were satisfied that they would remember it. When the participant was ready to continue, all three notifications would be delivered again in a random order. To complete the training, the participant simply had to acknowledge the target notification and ignore the two distractor notifications. If the participant failed the training then they would be returned to the screen shown in Figure 4.6 and asked to repeat the training segment with the same target notification.

When the training was complete participants would play 4 games of Concentration, each with a 60-second time limit. A total of 12 notifications would be delivered over the 4 games; 6 target notifications and 6 distractor notifications, as specified in Section 4.2.4. At the end of each game the participant would press a button to start the next game, giving an opportunity to rest if necessary.

After completing the control condition and each experimental condition the NASA-TLX survey was administered (provided in Appendix A). While the participant carried out this task the hardware used in that condition would be cleared away and any hardware required for the following condition would be set up.

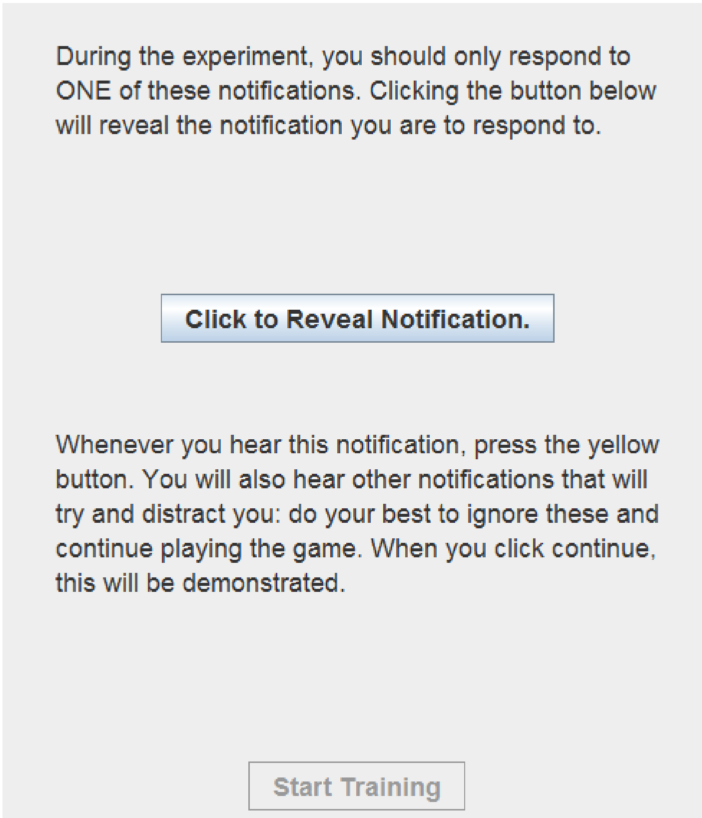


Figure 4.6: The screen shown to participants to introduce the target notification. The greyed-out button labelled 'Start Training' would become functional after the participant had received the target notification once, but participants were free to re-deliver the notification as many times as they wanted.

After all the conditions were completed the participant was debriefed and an informal exit interview was carried out. The experiment required around 60-90 minutes and participants were paid £10 each.

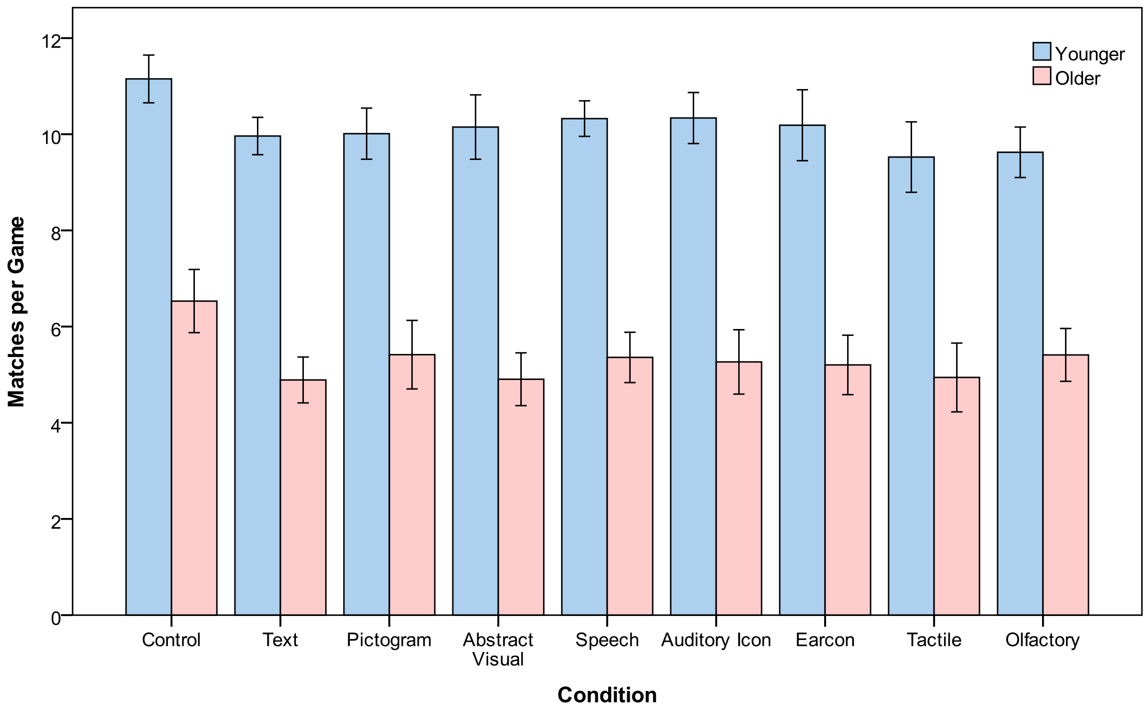


Figure 4.7: Graph showing the number of cards matched in a game by modality and age. Cards Matched ranges from 0-12. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

4.6 Results

This section presents the results of the study, organised by the hypotheses presented in Section 4.3.

4.6.1 Hypothesis 1.1 (Performance)

Hypothesis 1.1 evaluated the effects of age and modality on performance, measured by the number of cards matched per game. It is split into three sub-hypotheses which examine modality, age and any interactions between them. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using modality as the repeated-measures variable and age as the between-groups variable. An overview of the procedure used when carrying out GLM tests is provided in Appendix D.1. Mauchly's test was not significant ($\chi^2(35) = 45.7,\ p = .11$), so the assumption of sphericity was upheld and no correction was applied to the results of the test. The resulting data are shown in Figure 4.7.

Hypothesis 1.1.1 was "the modality used to deliver a notification will affect the number of cards matched". The model showed that modality had a significant main effect on the number of cards matched in a game ($F(8,\ 272) = 4.71,\ p < .001,\ \eta^2 = .12$). This is similar to the results found in the first study. However, *post hoc* pairwise comparisons (provided in Appendix B, Table B.1) revealed that the only performance differences lay between the control condition and the experimental conditions. Therefore, the different modalities had no significant effect on game performance; all appeared to have disrupted game performance to a similar level. The evidence in this case does not support the hypothesis.

Hypothesis 1.1.2 was "the age of the participant will affect the number of cards matched". As can clearly be shown in Figure 4.7, age had a strong main effect on game performance ($F(1,\ 34) = 85.9,\ p < .000,\ \eta^2 = .72$). This supports the hypothesis, although the effect was much stronger than anticipated.

Hypothesis 1.1.3 was "there will be no interaction between modality and age on the number of cards matched". The model found no interaction effect of modality and age on game performance ($F(8,\ 272) = 0.61,\ p = .77$), supporting the hypothesis.

These results suggest that the biggest factor influencing performance was age. While the additional workload of the secondary task caused a significant performance drop in the experimental conditions, by far the most significant effect was that of age. The implications of this result are discussed in Section 4.7.

4.6.2 Hypothesis 1.2 (Activity Rate)

Hypothesis 1.2 evaluated the effects of age and modality on activity rate, measured by the number of clicks made per second. It is split into three sub-hypotheses which examine modality, age, and any interactions between age and modality. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using modality as the repeated-measures variable and age as the between-groups variable. Only data gathered after target notifications were used here; data gathered after distractor notifications were not included in the model (but are evaluated in Section 4.6.11). Mauchly's test was significant ($\chi^2(35) = 54.51,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .68$). The data are shown in Figure 4.8.

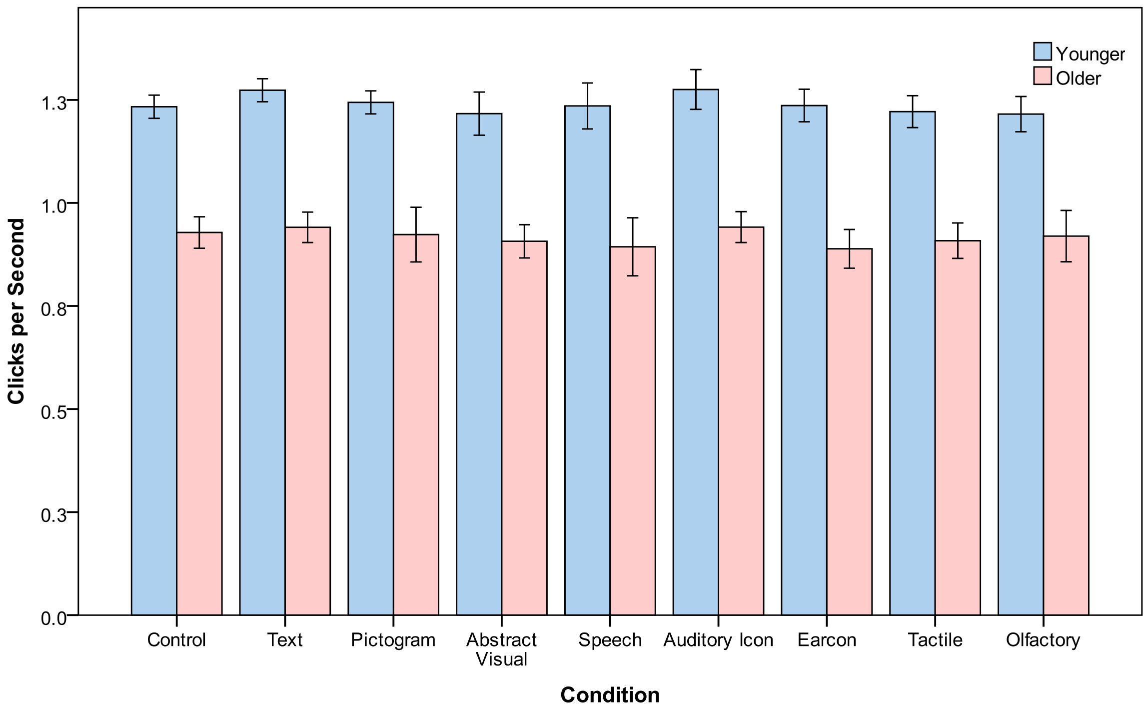


Figure 4.8: Graph showing the activity rate by modality and age. Activity rate is measured as the number of clicks per second. This graph does *not* include data gathered after distractor notifications (see Figure 4.14). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Hypothesis 1.2.1 was "the modality used to deliver a notification will affect the click rate". The model found no main effect of modality on activity rate ($F(5.41,\ 184.07) = 1.31,\ p = .23$). This does not support the hypothesis, and as it was not significant, no *post hoc* tests were carried out.

Hypothesis 1.2.2 was "the age of the participant will affect the click rate". Much like performance, age had a significant main effect on activity rate ($F(1,\ 34) = 37.1,\ p < .000,\ \eta^2 = .52$), which supports the hypothesis.

Hypothesis 1.2.3 was "there will be no interaction between modality and age on the click rate". The model showed no interaction effect of modality and age on the click rate ($F(5.41,\ 184.07) = 0.29,\ p = .93$), which supports the hypothesis.

Much like with performance, age was again the primary factor affecting activity rate. Modality was shown to have no significant effect on activity rate, which is particularly interesting given the model included a control condition with no secondary task. This result will be discussed in more detail in Section 4.7.

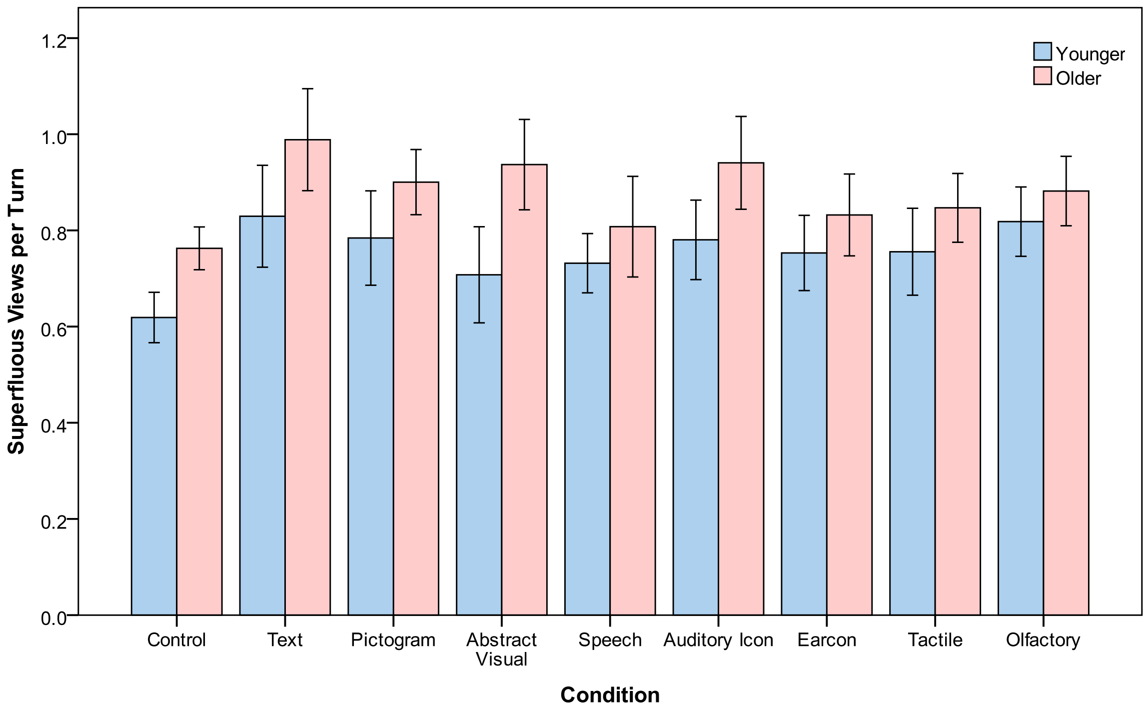


Figure 4.9: Graph showing the error rate by modality and age. Error rate was measured by the number of Superfluous Views per Turn (see Section 3.2.5), which ranges from 0-2. This graph does *not* include data gathered after distractor notifications (see Figure 4.15). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

4.6.3 Hypothesis 1.3 (Error Rate)

Hypothesis 1.3 tested the effects of age and modality on the error rate, measured by the number of superfluous views per turn (see Section 3.2.5). It is split into three sub-hypotheses which examine modality, age and any interactions between age and modality. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using modality as the repeated-measures variable and age as the between-groups variable. Only data gathered after target notifications were used here; data gathered after distractor notifications was not included in the model (but are evaluated in Section 4.6.10). Mauchly's test was not significant ($\chi^2(35) = 41.32,\ p = .22$) so sphericity was assumed for the model. The data are shown in Figure 4.9.

Hypothesis 1.3.1 was "the modality used to deliver a notification will not affect the number of superfluous views per turn". The model showed that modality did have a main effect on error rate ($F(8,\ 272) = 4.21,\ p < .001,\ \eta^2 = .11$). *Post hoc* pairwise comparisons, provided in Table B.2, show that the only significant differences were between the control condition and the experimental conditions. This result supports the hypothesis, as the different modalities did not produce significantly different error rates in the primary task.

Hypothesis 1.3.2 was "the age of the participant will affect the number of superfluous views per turn". The model showed that unlike other disruption measures, age did not have a significant main effect on error rate ($F(1,\ 34) = 3.28,\ p = .08$).

Hypothesis 1.3.3 was "there will be no interaction between modality and age on the number of superfluous views per turn". The model supports this hypothesis as no interaction effect was found of modality and age on error rate ($F(8,\ 272) = 0.29,\ p = .62$).

This result was unexpected as it shows that the error rate was generally the same despite the effects of age and modality; the other results seem to suggest that the poor performance of older players was actually a result of a generally lower activity rate. A more thorough interpretation of this finding is provided in Section 4.7.

4.6.4 Hypothesis 2.1 (Response Accuracy)

Hypothesis 2.1 tested the effects of age and modality on response accuracy. It is split into three sub-hypotheses which examine modality, age, and any interactions between age and modality. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using modality as the repeated-measures variable and age as the between-groups variable. Mauchly's test was found to be significant ($\chi^2(27) = 97.79,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .63$). The data are shown in Figure 4.10.

Hypothesis 2.1.1 was "response accuracy will be affected by the modality used to deliver the notifications". A main effect of modality on response accuracy was found ($F(4.44,\ 150.85) = 13.47,\ p < .001,\ \eta^2 = .28$). *Post hoc* pairwise comparisons shown in Table B.3 reveal that, much like the results from the first study, there were two groups: one containing the 6 visual and audio conditions and the other containing the tactile and olfactory conditions. The evidence in this case supports the hypothesis that modality would affect response accuracy.

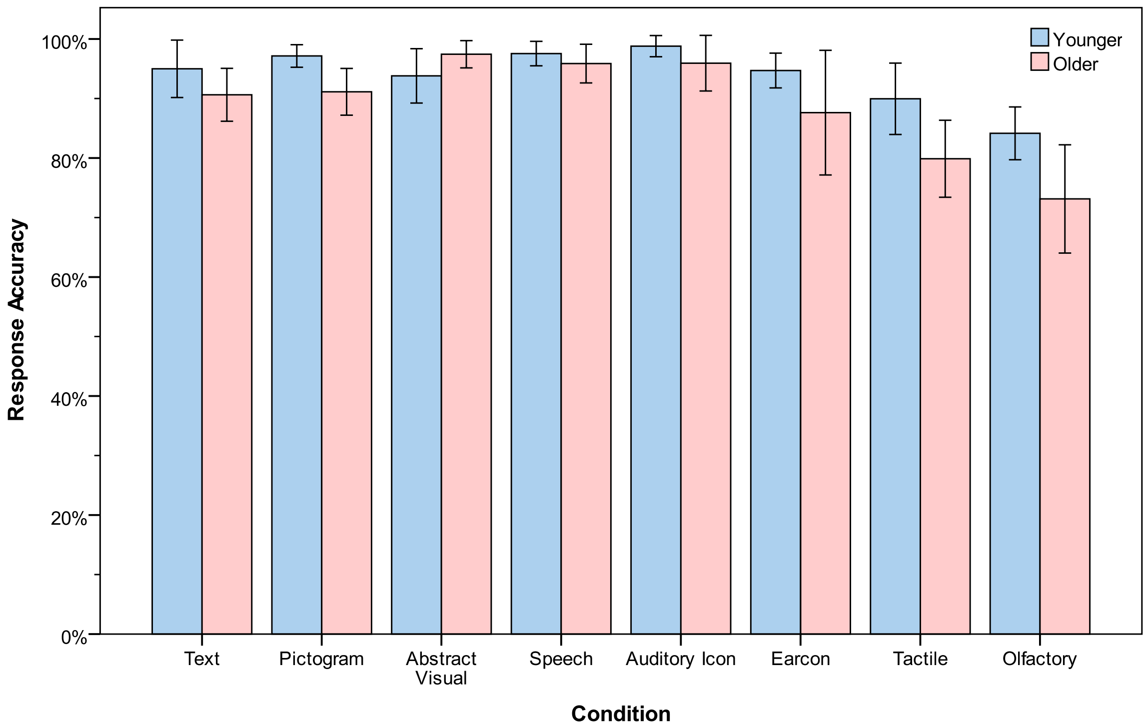


Figure 4.10: Graph showing the accuracy of responses to notifications by modality and age. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Hypothesis 2.1.2 was "response accuracy will not be affected by age". The model in this case shows that age did have an effect on the notification response accuracy ($F(1, 34) = 9.2,\ p < .01,\ \eta^2 = .21$), which does not support the hypothesis.

Hypothesis 2.1.3 was "age and modality will not interact to affect response accuracy". The model found that any interaction between modality and age on response accuracy did not reach significance ($F(4.44, 150.85) = 1.8,\ p = .13$), supporting the hypothesis.

These results showed that the older user group, contrary to expectations, did not match the performance of younger users in the secondary task. As expected, modality affects response accuracy, but Figure 4.10 also seems to suggest there may be an interaction effect; age seems to have a more pronounced effect in the earcon, tactile and olfactory conditions. However, the model showed that these effects were not significant. This result will be explored in Section 4.7.

4.6.5 Hypothesis 2.2 (Processing Time)

Hypothesis 2.2 tested the effects of age and modality on processing time, which was measured by the longest pause in activity after a notification had been delivered. It

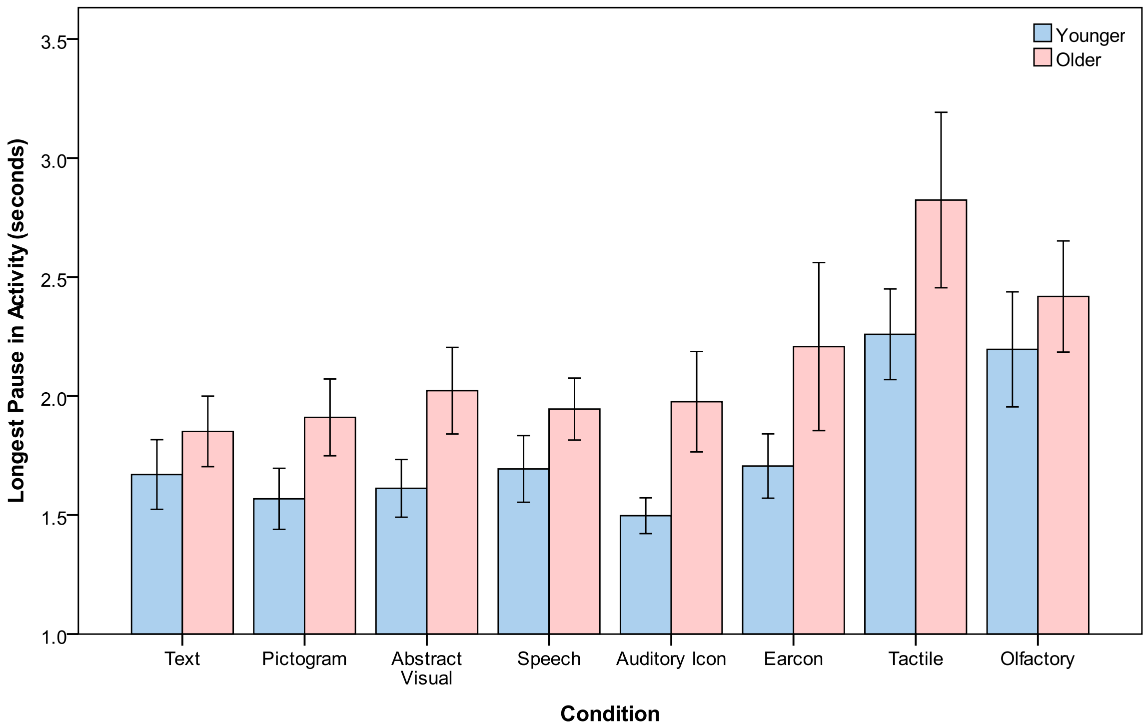


Figure 4.11: Graph showing the time taken to process a target notification, measured by the longest pause (in seconds). This graph does *not* include data gathered after distractor notifications (see Figure 4.16). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

is split into three sub-hypotheses which examine modality, age, and any interactions between age and modality. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using modality as the repeated-measures variable and age as the between-groups variable. Only data gathered after target notifications were used here; data gathered after distractor notifications was not included in the model (but are evaluated in Section 4.6.12). Mauchly's test was found to be significant ($\chi^2(27) = 84.4,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .61$). The data are shown in Figure 4.11.

Hypothesis 2.2.1 was "processing time will be affected by the modality used to deliver the notifications". The model showed a main effect of modality on longest pause ($F(4.25,\ 144.49) = 18.91,\ p < .001,\ \eta^2 = .36$). *Post hoc* pairwise comparisons, included in Table B.4, revealed that no differences between the visual and audio conditions, but significant differences between that group and the tactile and olfactory conditions. This result supports the hypothesis.

Hypothesis 2.2.2 was "processing time will be affected by age". The model showed that age had a significant effect on processing time ($F(1, 34) = 7.59,\ p < .01,\ \eta^2 = .18$), which supports the hypothesis.

Hypothesis 2.2.3 was "age and modality will not interact to affect processing time". The model showed no interaction effect of modality and age on the processing time ($F(4.25, 144.49) = 0.1.05,\ p = .39$), which supports the hypothesis.

These results show that modality played a considerable role in the time taken to process an notification, with the olfactory and tactile modalities performing the worst. Surprisingly, the impact of age is quite small compared the large disruptive effects observed. This funding is discussed further in Section 4.7.

4.6.6 Hypothesis 2.3 (Delivery Time)

Hypothesis 2.3 tested the effects of age and modality on delivery time, calculated by measuring the distance from the start of notification delivery to the start of the longest pause. It is split into three sub-hypotheses which examine modality, age, and any interactions between age and modality. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using modality as the repeated-measures variable and age as the between-groups variable. Only data gathered after target notifications were used here; data gathered after distractor notifications was not included in the model. Mauchly's test was found to be significant ($\chi^2(27) = 245.33,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .23$). The data are shown in Figure 4.12.

Hypothesis 2.3.1 was "delivery time will be affected by the modality used to deliver the notifications". The model confirmed that modality had a main effect on delivery time ($F(1.62, 55.15) = 114.14,\ p < .001,\ \eta^2 = .77$). *Post hoc* pairwise comparisons were carried out and are included in Appendix B.1, Table B.5. As can be seen in Figure 4.12, there is little variance within modalities with the exception of the olfactory condition. The low variance resulted in the tactile and olfactory conditions being significantly different from the other conditions, even though for tactile the real-world difference was very small. This result supports the hypothesis.

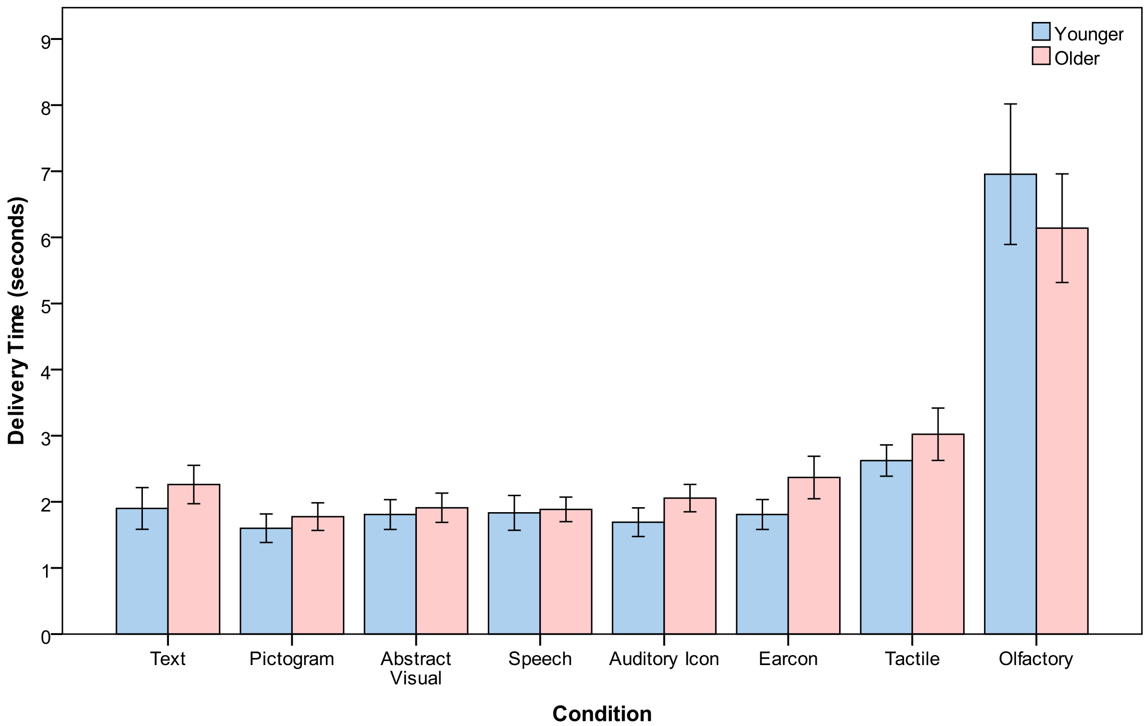


Figure 4.12: Graph showing the estimated delivery time for notifications. Deliver time is the time in seconds between the start of notification delivery and the start of the longest pause. This graph does *not* include data gathered after distractor notifications. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Hypothesis 2.3.2 was "delivery time will not be affected by age". As expected, the model shows no main effect of age on the time taken to deliver a notification ($F(1, 34) = .73,\ p = .4$).

Hypothesis 2.3.3 was "age and modality will not interact to affect delivery time". The model found that any interaction between modality and age on response accuracy did not reach significance ($F(1.62, 55.15) = 1.94,\ p = .16$), supporting the hypothesis.

These results show that the deliver time was generally consistent, with no real effect of age. Most of the modalities were delivered in the same amount of time, although the olfactory condition took much longer than the other methods. This is discussed further in Section 4.7.

4.6.7 Hypothesis 3.1 (Workload – Younger)

Hypothesis 3.1 was defined in Section 4.3.3 as "for younger participants, subjective workload will be affected by the modality used to deliver notifications". As discussed

Table 4.2: Analysis of repeated-measures TLX factors for the younger participant group.

TLX	χ^2	Kendall's W	N	df	p
OW	41.87	0.26	20	8	$<.001^{***}$
MD	32.99	0.21	20	8	$<.001^{***}$
PD	22.66	0.14	20	8	0.004^{**}
TD	23.87	0.15	20	8	0.002^{**}
OP	28.02	0.18	20	8	$<.001^{***}$
EF	31.40	0.20	20	8	$<.001^{***}$
FR	29.41	0.18	20	8	$<.001^{***}$

Note: Table shows Freidman's two-way ANOVA calculated using SPSS. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$. *Post hoc* tests are provided in Appendix B.

in Section 3.2.5, Workload (WL) was measured using the NASA-TLX assessment, which uses 6 independent components to calculate an overall workload. While overall workload is the primary measure, the individual elements also provide an insight into workload differences. The overall workload might disguise significant differences between components (*e.g.*, one condition might produce a significantly higher Mental Demand (MD) which could be hidden by the aggregated overall workload score. The overall workload is represented by the total of 6 workload component measures: MD, Physical Demand (PD), Temporal Demand (TD), Overall Performance (OP), Effort (EF) and Frustration (FR). The overall workload and its composition is shown in Figure 4.13a.

Table 4.2 shows the result of a set of Friedman's ANOVAs testing for an effect of modality on the NASA-TLX scores. This table shows that for younger participants, all 6 primary measures and the overall workload itself were affected by the modality with a high degree of significance. To understand which modalities produced the higher workload levels, *post hoc* tests were carried out which are included in Appendix B. The *post hoc* can be summarised as follows:

WL The control condition was found to be different from the text, pictogram, earcon, tactile and olfactory conditions. Between experimental conditions, only two significant differences were found: speech-olfactory and auditory icon-olfactory (Table B.1b).

MD Significant differences only existed between the control condition and the text, pictogram, tactile and olfactory conditions (Table B.2b).

PD Significant differences only existed between the control condition and the tactile and olfactory conditions (Table B.3b).

TD Significant differences only existed between the text and control conditions (Table B.4b).

OP Significant differences existed between the control condition and the text, pictogram, tactile and olfactory conditions (Table B.5b).

EF Significant differences were found between the control condition and the text and and olfactory conditions. The olfactory condition was also found to be significantly different from speech, auditory icon and earcon conditions (Table B.6b).

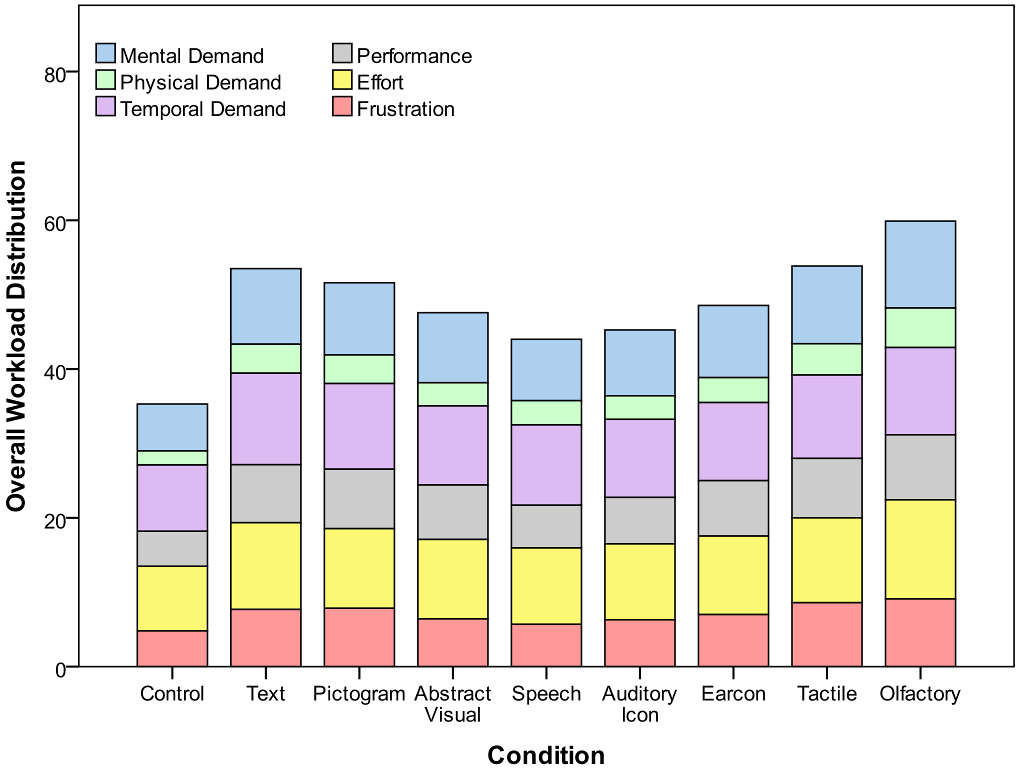
FR Significant differences only existed between the control condition and the pictogram, tactile and olfactory conditions (Table B.7b).

These results are somewhat similar to the findings of Chapter 3 (see Section 3.6.7) in that most of the significant differences lie between the control condition and the experimental conditions. There are very few significant differences between the experimental conditions, suggesting that the modality used to deliver a notification does not affect subjective workload. This was surprising given the previous results, which suggested a higher workload associated with tactile and olfactory notifications. This is discussed in further in Section 4.7.

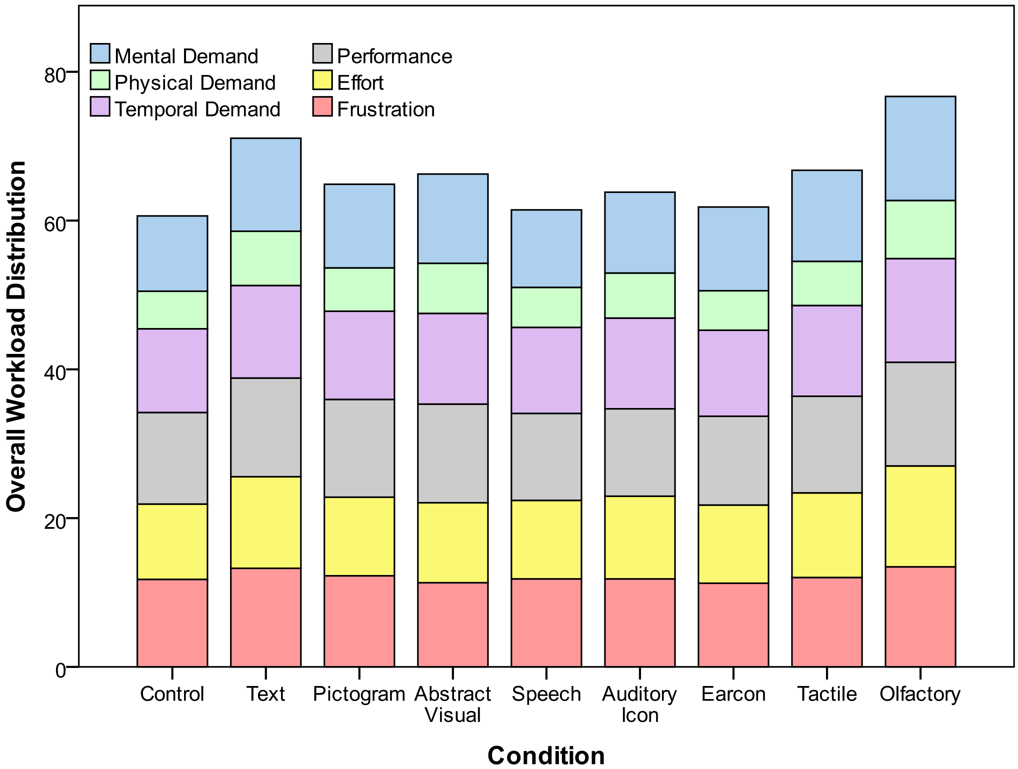
4.6.8 Hypothesis 3.2 (Workload – Older)

Hypothesis 3.2 was defined in Section 4.3.3 as "for older participants, subjective workload will be affected by the modality used to deliver notifications". The overall workload and its composition for the older participant group is shown in Figure 4.13b. Table 4.3 shows the result of a set of Friedman's ANOVAs testing for an effect of modality on the NASA-TLX scores. The table shows that the modality affected overall workload along with all workload components with the exception of frustration. As with younger participants, *post hoc* tests were carried out to identify where the significant differences lie, which are included in Appendix B. The main findings are as follows:

WL Significant differences were only found between the control condition and the text and olfactory conditions (Table B.1b).



(a) Younger Participant Group



(b) Older Participant Group

Figure 4.13: Graph showing the composition of the NASA-TLX Overall Workload metric. Graphs for the individual NASA-TLX components are included in Appendix B (Figure B.1 provides a more direct comparison of workloads between the older and younger participants).

Table 4.3: Analysis of repeated-measures TLX factors for the older participant group.

TLX	χ^2	Kendall's W	N	df	p
OW	23.64	0.19	16	8	0.003**
MD	22.63	0.18	16	8	0.004**
PD	25.12	0.20	16	8	0.001**
TD	21.46	0.17	16	8	0.006**
OP	16.43	0.13	16	8	0.037*
EF	19.45	0.15	16	8	0.013*
FR	11.82	0.09	16	8	0.160

Note: Table shows Freidman's two-way ANOVA calculated using SPSS. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$. *Post hoc* tests are provided in Appendix B.

MD Significant differences only existed between the olfactory condition and the control and speech conditions (Table B.2b).

PD Significant differences were found only between the text condition and the control, speech and earcon conditions (Table B.3b).

TD Significant differences existed only between the olfactory condition and the control, pictogram, speech and earcon conditions (Table B.4b).

OP Significant differences existed between the control condition and the text, pictogram, tactile and olfactory conditions (Table B.5b).

EF Significant differences were found between the control condition and the olfactory condition (Table B.6b).

FR No *post hoc* tests were run for frustration, as Table B.6b shows that the condition did not affect frustration.

The results for older participants tend to exhibit similar patterns as the results for younger participants, in that most of the differences existed between the control condition and experimental conditions. Therefore, as with the younger participants the evidence does not support the hypothesis in this case. There are fewer *post hoc* significant differences for the older group compared to the younger group, which may be due to the smaller participant group and the large number of conditions; *i.e.* there is a chance that this conclusion is a type II error. This is discussed in Section 4.7.

Table 4.4: Between-groups tests for the workload components.

TLX	U	Wilcoxon W	N	Z	p	p_{adj}	r
OW	240.50	376.50	36	2.563	0.009	0.005^{**}	0.43
MD	208.00	344.00	36	1.529	0.132	0.066	0.25
PD	222.50	358.50	36	1.991	0.046	0.023^{*}	0.33
TD	179.50	315.50	36	0.535	0.539	0.270	0.09
OP	280.00	416.00	36	3.821	0.000	$<.001^{***}$	0.64
EF	174.00	310.00	36	0.446	0.671	0.336	0.07
FR	258.00	394.00	36	3.121	0.001	$<.001^{***}$	0.52

Note: Table shows Mann-Whitney's U test and Wilcoxon's W. The tests were carried out using SPSS which produces two-tailed significance values (listed as p), but as the hypotheses were one-tailed the p values were adjusted to suit (listed as p_{adj}). Effect sizes are shown under r, calculated using the method suggested by Rosenthal [177] which is $r = \frac{Z}{\sqrt{N}}$. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

4.6.9 Hypothesis 3.3 (Workload – Between Groups)

Hypothesis 3.3 was defined in Section 4.3.3 as "older participants will report higher subjective workloads than younger participants". This was testing by taking the mean NASA-TLX scores for each participant and comparing the means using the independent-samples Mann-Whitney U test. The results of this process are shown in Table 4.4, and graphs comparing the workload components are provided in Appendix C.

The results showed showed a significant effect of age on the overall workload, physical demand, performance and frustration. The NASA-TLX performance effect was particularly significant due to the size of the effect ($r = .64$) and the magnitude of the difference in means (visible in Figure B.5). The overall workload is shown in Figure C.1a, which demonstrates that the differences were in the predicted direction. The evidence in this case supports the hypothesis that older participants would report higher workload scores. This finding will be discussed further in Section 4.7.

4.6.10 Hypothesis 4.1 (Distraction & Activity Rate)

Hypothesis 4.1 tested the effects of notification relevance on activity rate post-notification. It is split into three sub-hypotheses which examined possible interaction effects with modality and age. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using notification relevance and modality as repeated-measures variables and age as a between-groups variable. The data are shown in Figure 4.14.

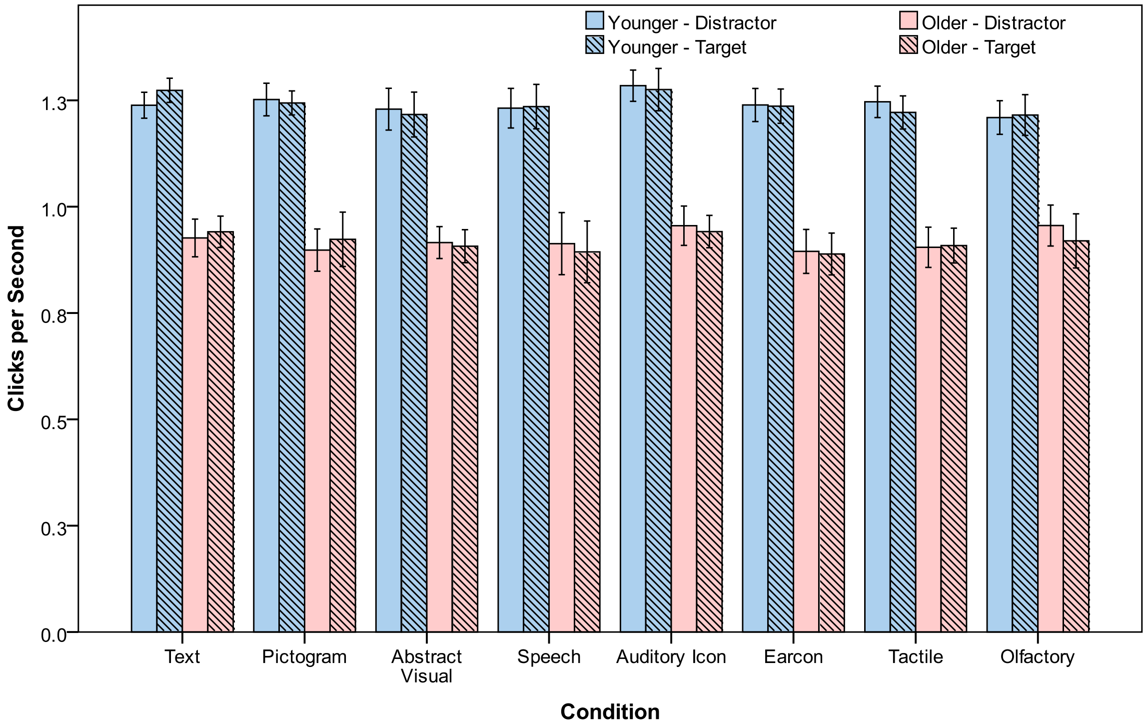


Figure 4.14: Graph showing how notification relevance, age and modality affected activity rate. Error bars show 95% confidence intervals.

Hypothesis 4.1.1 was "notification relevance will not affect click rate". Mauchly's test was not significant ($\chi^2(27) = 37.32,\ p = .09$), so the assumption of sphericity was held. The model showed no main effect of notification relevance on activity rate ($F(1,\ 34) = .44,\ p = .51$), supporting the hypothesis.

Hypothesis 4.1.2 was "notification relevance will not interact with modality to affect click rate". Mauchly's test was not significant ($\chi^2(27) = 28.34,\ p = .4$), so the assumption of sphericity was held. The model showed no main effect of notification relevance on activity rate ($F(7,\ 238) = 1.3,\ p = .25$), supporting the hypothesis.

Hypothesis 4.1.3 was "notification relevance will not interact with age to affect click rate". The model showed that relevance and age did no produce an interaction effect on activity rate ($F(1,\ 34) = .12,\ p = .74$), which supports the hypothesis.

These results show that while age had a significant effect on activity rate (as shown in Section 4.6.2), there was little difference in post-notification activity rate between target and distractor notifications and no interactions existed. This result will be explored in Section 4.7.

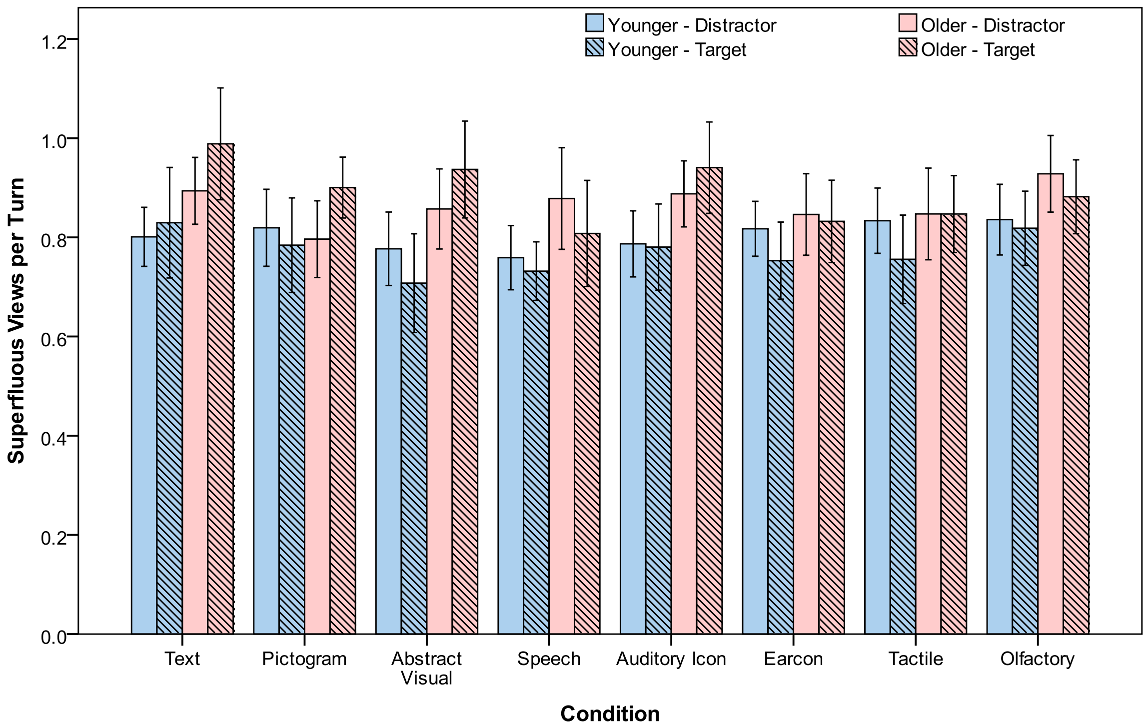


Figure 4.15: Graph showing how notification relevance, age and modality affected error rate (Superfluous Views per Turn). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

4.6.11 Hypothesis 4.2 (Distraction & Error Rate)

Hypothesis 4.2 tested the effects of notification relevance on error rate post-notification, measured by the number of superfluous views per turn. It is split into three sub-hypotheses which examined possible interaction effects with modality and age. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using notification relevance and modality as repeated-measures variables and age as a between-groups variable. The data are shown in Figure 4.15.

Hypothesis 4.2.1 was "notification relevance will not affect the number of superfluous views per turn". Mauchly's test was not significant ($\chi^2(27) = 27.94,\ p = .42$), so the assumption of sphericity was held. The model showed no main effect of relevance on error rate ($F(1, 34) = .19,\ p = .67$), supporting the hypothesis.

Hypothesis 4.2.2 was "notification relevance will not interact with modality to affect the number of superfluous views per turn". Mauchly's test was not significant ($\chi^2(27) = 13.73,\ p = .98$), so the assumption of sphericity was held. No interaction effect was found from modality and relevance on error rate ($F(7, 238) = .99,\ p = .44$), which supports the hypothesis.

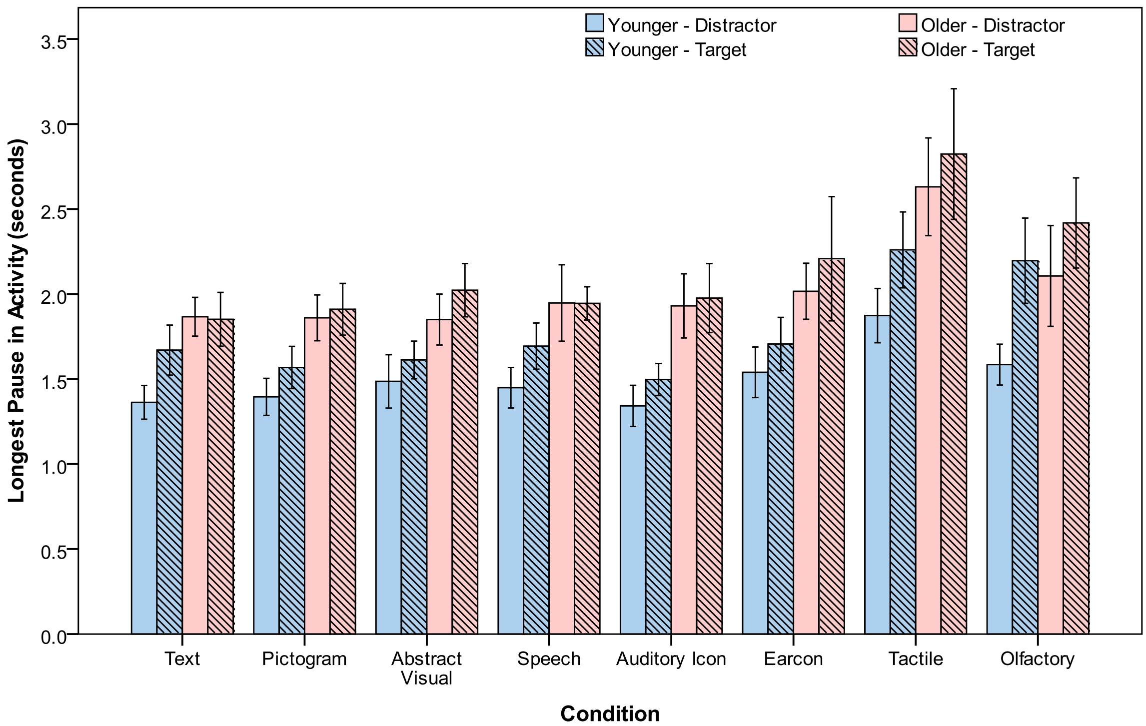


Figure 4.16: Graph showing how notification relevance, age and modality affected processing time (longest pause). Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

Hypothesis 4.2.3 was "notification relevance will not interact with age to affect the number of superfluous views per turn". The model showed that error rate was affected by an interaction between relevance and age ($F(1, 34) = 8.65,\ p < .01,\ \eta^2 = .2$), which does not support the hypothesis.

The results showed one surprising finding: age and notification relevance seemed to interact. The results seem to suggest that younger users had a higher error rate after distractions, while older users had a higher error rate after target notifications. This will be explored in more detail in Section 4.7.

4.6.12 Hypothesis 4.3 (Distraction & Processing Time)

Hypothesis 4.3 tested the effects of notification relevance on processing time, measured by the longest pause. It is split into three sub-hypotheses which examined possible interaction effects with modality and age. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using notification relevance and modality as repeated-measures variables and age as a between-groups variable. The data are shown in Figure 4.16.

Hypothesis 4.3.1 was "distractor notifications will produce a shorter 'longest pause' than target notifications." Mauchly's test was found to be significant ($\chi^2(27) = 57.67,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .65$). Notification relevance was found to have a significant main effect on longest pause ($F(1,\ 34) = 31.52,\ p < .001,\ \eta^2 = .48$), supporting the hypothesis.

Hypothesis 4.3.2 was "notification relevance and modality will interact to affect longest pause". Mauchly's test was found to be significant ($\chi^2(27) = 84.11,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .56$). The model showed an interaction effect did exist ($F(3.93,\ 133.52) = 2.65,\ p < .05,\ \eta^2 = .07$), which supports the hypothesis.

Hypothesis 4.3.3 was "notification relevance will not interact with age to affect longest pause". However, the model showed that age and relevance did create an interaction effect on longest pause ($F(1,\ 34) = 4.84,\ p < .05,\ \eta^2 = .13$), which does not support the hypothesis.

This result shows that the processing time, defined as the longest pause in activity, is actually affected by a number of factors including age, modality and relevance. While these factors created individual interactions, there was not a combined interaction effect between age, modality and relevance (($F(3.93,\ 133.52) = .84,\ p = .5$). This finding is discussed in Section 4.7.

4.6.13 Hypothesis 4.4 (Distraction & Delivery Time)

Hypothesis 4.4 tested the effects of notification relevance on delivery time. It is split into three sub-hypotheses which examined possible interaction effects with modality and age. The hypotheses were tested by constructing a mixed-models general linear model (GLM) using notification relevance and modality as repeated-measures variables and age as a between-groups variable. The data are shown in Figure 4.17.

Hypothesis 4.4.1 was "notification relevance will not significantly affect delivery time". Mauchly's test was found to be significant ($\chi^2(27) = 190.25,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .28$). Notification relevance was found to

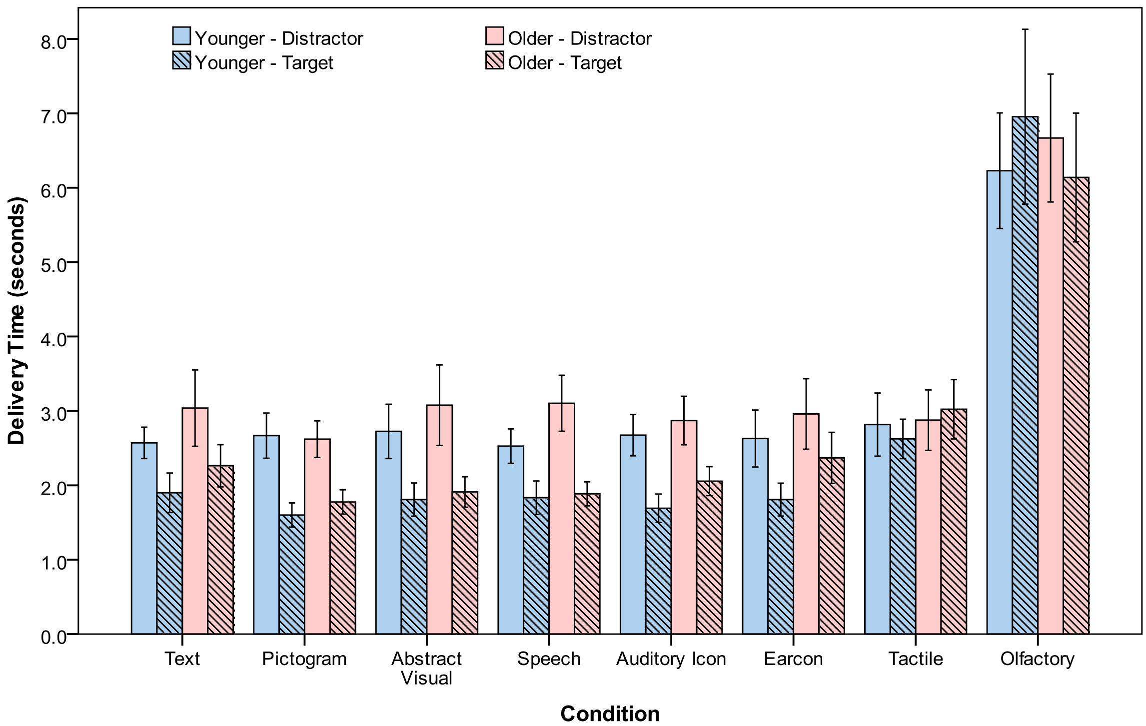


Figure 4.17: Graph showing how notification relevance, age and modality affected delivery time. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

have a significant main effect on longest pause ($F(1, 34) = 54.76,\ p < .001,\ \eta^2 = .62$), which does not support the hypothesis.

Hypothesis 4.4.2 was "notification relevance and modality will not interact to affect delivery time". Mauchly's test was found to be significant ($\chi^2(27) = 105.15,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .43$). Notification relevance was found to interact with modality to affect delivery time ($F(3, 102.07) = 8.07,\ p < .01,\ \eta^2 = .14$), which does not support the hypothesis.

Hypothesis 4.4.3 was "notification relevance will not interact with age to affect delivery time". As expected, no such interaction was found ($F(1, 34) = .7,\ p = .41$).

These results are interesting because it seems to suggest that participants responded more quickly to useful information than they did distractors. As can be seen in Figure 4.17 however, this seemed to be more prevalent for some modalities than for others. This finding will be discussed in the following Section.

4.7 Discussion

In this section the results of the experiment are discussed. Section 4.7.1 discusses disruption, followed by Section 4.7.2 which discusses the effectiveness of the notifications. Section 4.7.3 explores the subjective findings of the experiment. Section 4.7.4 explores how the presence of distractor notifications affected the experimental participants.

4.7.1 Disruption

The first hypothesis, presented in Section 4.6.1, showed a significant effect of modality on performance. However, the pairwise comparisons (provided in Table B.1) showed that the only real differences lay between the control and experimental conditions. This is different from the findings from the first experiment, which found that the olfactory and tactile conditions performed more poorly than the visual and audio conditions.

It is possible that this difference is actually a type II error due to the increased number of conditions and the Sidak correction. Comparing the mean cards matched between the first and second studies could give an indication of this. However, this study has shown that age significantly affects cards matched; as the first study included a range of age groups, it is not possible to make a reliable comparison between the two studies. Therefore, the chance that this difference is due to a type II error cannot be ruled out in this way.

One theory suggested to explain the lower performance in the tactile and olfactory conditions in the first study (see Section 3.7) was that those modalities took longer to respond to, which resulted in a lower activity rate and thus lower performance. The modality itself was not responsible for the disruption and lower performance; in fact, the modality of the interruption in the first study did not seem to affect the process of stopping and resuming a task. The evidence from this study seems to confirm that this theory is correct; the difference in the overall performance was not significant between modalities, nor was there a significant difference in error rate (although again, given the large number of conditions there is a higher chance of committing a type II error).

One of the most striking results from the study was a considerable difference between the performance of the older and younger participants, as shown in Figure 4.7. As was shown in Section 4.6.1, the power of this effect was considerable ($\eta^2 = .72$). Given that performance is a dictated by activity rate and error rate, which is most responsible for the lower performance of older participants?

As is clearly shown in Figure 4.8, older participants were much slower compared to younger participants. The size of this effect was also quite strong ($\eta^2 = .52$). Park [169] suggested that the most likely reason for this difference is a reduction in the speed of working memory. As shown in Section 4.6.3, any effect of age on the error rate did not reach significance. This suggests that the primary source of lower performance with older participants was related to a lower activity rate.

Arroyo *et al.* [9] had tested disruption subjectively, but found no significant differences in any objective measures. This study agrees with their results, suggesting that there is no relationship between notification modality and error rate. Cellier & Eyrolle [43] suggested an interrupted person takes a 'mental snapshot' of the task at hand when interrupted and then uses that information to resume the task later. In Section 3.7 it was suggested that the modality of the notification does not affect this process, and the results of this study provide more evidence that this theory is valid. While the modality does appear to affect the player's behaviour in the game, there is no evidence of an effect on performance, error rate or activity rate that can be directly attributed to modality.

4.7.2 Effectiveness

As expected, hypothesis 2.1 (Section 4.6.4) showed that notification response accuracy was heavily affected by modality. Much like the first study, the olfactory and tactile conditions did not match the visual and audio conditions in terms of performance. However, the results for this study showed higher levels of performance than the results of the first study. An interaction effects seems to suggest that the performance of the older participants was similar to that of younger participants for the visual and audio conditions, but performance was lower in the tactile and olfactory conditions. Despite this, the worst performance was by the older participants in the olfactory condition, which was still quite high at 73%. It is likely that the performance differences between the two experiments are due to the less demanding secondary task.

The delivery time measures showed no effect of age, as expected, but did show an effect of modality. All the visual and audio conditions were not significantly different, but were shown to be significantly quicker than tactile and olfactory notifications. The tactile delay is interesting because the hardware for delivering tactile notifications should match the speed of the audio conditions. Although the differences were found to be significant, the actual difference between the tactile condition and the visual/audio conditions was very small at around 0.9 seconds. As the tactile technology is similar to the audio technology, this effect cannot be a result of latency in the hardware; it may be evidence of a neurological or somatosensory delay in becoming aware of tactile information.

The processing time, shown in Figure 4.11, displayed a similar effect. Modality had an effect, but *post hoc* pairwise comparisons revealed two groups: one visual/audio, the other tactile/olfactory. The average processing time for visual/audio notifications was around 1.8 seconds, with 2.5 seconds for tactile and 2.3 seconds for olfactory. This result seems to confirm the existence of some kind of 'tactile lag' as suggested in the first study (see Section 3.7). The lack of significant differences between the tactile and olfactory conditions suggest a similar effect may also exist for olfactory. If so, then the 'tactile lag' may simply be evidence that the visual and auditory systems are more powerful than the tactile and olfactory systems at receiving and processing data. However, many participants verbally expressed that they had forgotten the button/notification associations at the start of the tactile condition, which did not occur with any of the other notifications. Despite this, tactile response accuracy was still quite high (Younger = 90%, Older = 80%).

An alternative theory is that of training or familiarity. Anecdotal evidence suggests that disruption could be greater for modalities which the user is unfamiliar with [9]; this is backed up by research showing that disruption effects can be reduced through experience or training [40, 92]. This study did not measure familiarity, but the difference in effectiveness between visual/auditory conditions and the tactile and olfactory conditions could be explained by this effect. Given the ubiquity of visual and audio devices it seems reasonable to assume that participants were more 'familiar' with visual and auditory notifications. Although tactile devices are now common in mobile phones, they generally produce simple vibrations instead of structured tactile messages. Familiarity might provide an increased ability to intercept and process these notifications, but further experiments would be needed to evaluate this.

4.7.3 Subjective Workload

The subjective workload ratings for both the older and younger users suggest that the workload didn't vary significantly between modalities. There is a validity issue with the experiment introduced by the number of conditions and the number of participants. When correcting for the number of conditions, a Bonferroni correction would be inappropriate as it is highly conservative (even when faced with smaller conditions) [63]. The *post hoc* tests were carried out using R, which allowed the

Tukey-Kramer method to be applied instead, which is much more appropriate in this case. While the results showed little significance between the experimental conditions, they did show that the control condition generally produced a lower workload level compared to the experimental conditions. As this reflects the findings presented in Chapter 3, it suggests that the results are somewhat reliable. Figure 4.13 confirms that the control workload was generally lower than the experimental condition workloads. It is possible that the cases where the control condition was *not* significantly different from the experimental conditions is a type II error caused by the familywise error correction.

The between-groups statistics shown in Section 4.6.9 revealed that the older participants rated their performance to be lower and their frustration to be higher, which had an effect on overall workload. Section 4.6.1 showed that the lower performance rating is justified. It is interesting that this effect was large enough to have a significant impact on overall workload. The higher frustration level is particularly important, as frustration is likely to be a factor in technology rejection. This underlines the necessity of carrying out studies such as this with both younger and older participants, as the younger user group in this case acts as a control of sorts. Mental demand, temporal demand and effort are not significantly different between the groups, but the comparison shows that physical demand, performance and effort are higher for older participants.

Ideally, it would have been preferable to build a non-parametric mixed-designs model to compare the groups and conditions together, which would also have revealed interactions (as was done for the parametric statistics). This approach was attempted, but was not successful. More information on this is provided in Appendix D.3.

In the post-test interviews, older participants reacted very differently to the different modalities than younger participants. Despite younger and older participants exhibiting similar performance in the olfactory condition, younger participants were much more positive about smell while almost all of the older participants expressed negative sentiments. Participants were generally positive about the other modalities. Many participants stated that the abstract visual method was highly salient, although participants expressed different opinions on whether this was a good thing or not. Unlike with younger participants, older participants were highly aware of how these methods could be employed to interact with people who have sensory impairments.

In Chapter 3, the NASA-TLX data appeared to be more closely linked to the primary task (playing the card matching game) than the secondary task (responding to notifications). If this is also true in these experiments, then the workload changes only appear to highlight the added difficulty of multitasking (demonstrated by the repeated-measures factors only showing significance between the control and experimental conditions) and the effects of age on game performance (demonstrated by the increased frustration and lower performance measures). In general, the evidence suggest that there is not a significant relationship between workload and modality, refuting the hypotheses in Section 4.3.3 and shedding some light on the third research question specified in Section 4.1.

4.7.4 Distraction

The results of the study showed that there was no effect of notification relevance on activity rate or error rate, which was in line with expectations. However, the results showed an interaction between age and relevance on error rate in Section 4.6.11. The effect of this was not very strong ($\eta^2 = .2$), but as can be seen in Figure 4.15, the results appear to show more errors being made after target visual notifications (*i.e.* text, pictograms and abstract-visual) compared to the other modalities. This could be cross-modal interference between the visual primary task and the visual notifications. Latorella [123] found similar visual-visual cross-modal interference, but noted that the size of the effect was very small. Latorella found that the strongest effect came from audio-audio tasks. Interestingly, this effect is not seen for younger participants, which may suggest that older people are more susceptible to the negative effects of cross-modal interference.

Distractor notifications were found to have a shorter processing time than target notifications, which was expected given that they did not require any action on the part of the user. An interaction effect was found for age and relevance on processing time, however; while younger participants uniformly processed distractors faster than targets, older participants took roughly the same amount of time to process targets and distractor notifications in the visual and audio modalities. Notification relevance also interacted with modality; in particular, the the tactile and olfactory notifications had a much larger difference in processing time compared to the other modalities. As can be seen in Figure 4.16, the additional overhead of target notifications was very small in most of the modalities. This shows that the distractor notifications required more time to process than might be expected, and cannot be dismissed. Age was already shown to have an effect on processing time in Section 4.6.5, however Figure 4.16 shows that the younger participants could recognise and acknowledge target notifications quicker than older participants could identify and ignore distractor notifications in every condition except olfactory.

Delivery time was not expected to show any changes with notification relevance, yet a strong main effect was found ($\eta^2 = .62$). As shown in Figure 4.17, the delivery time for target notifications was less than the delivery time for distractor notifications in all conditions excluding tactile and olfactory. An interaction effect was also found with modality, which confirmed that the tactile and olfactory conditions exhibited different properties from the visual and audio conditions. No interaction was found with age; both younger and older participants responded to notifications quicker if the information was important. This suggests that there is some mental process which takes place where person 'looks out' for important information in the environment, therefore making it more salient. This seems to take place in every condition with the exception of tactile (both groups) and olfactory (younger only).

The fourth research question aimed to answer questions surrounding the effects of distractions in this type of technology. The results suggest that distractions appear to produce the same effects as target notifications. The processing time of distractor notifications, while significantly shorter than target notifications, is still quite long. The results also seem to show that participants responded quicker to useful information, while less useful information.

4.7.5 Reflection

The 'longest pause' measurement is assumed to represent the real point at which the participant processes the notification. However, this measurement was created specifically for this experiment and has not been used in the context of Concentration before. This forms one of the primary shortcomings of the measurement, with the other main shortcoming being that any other pauses inside the measurement period could disrupt the measurements. Several measures were taken to ensure that this measurement was reliable. Firstly, the measurement periods were carefully selected so that measurement only occurred after notification delivery was initiated, which reduced the chance of accidentally picking up other breaks in activity. For target notifications, the vast majority of longest pause values were found to overlap with the time the button was pressed (with the few that did not overlap generally being after the first action taken immediately after the button press). This check could not be carried out for the distractor notifications, so there is an assumption that the longest pause in that case represents the time spent mentally processing the notification. Finally, the resulting values have a fairly small amount of variance; most of the conditions showed similar pause lengths with the interesting exception of Tactile, which matches well with observations made after the first experiment. While the longest pause measure is not well-grounded in literature concerning Concentration as a test, Chapter 3 made it clear that existing measures were insufficient for the needs of this experiment, and the data suggests that longest pause was an appropriate and reliable measurement.

Figure 3.13 shows that the educational background and computer experience of both age groups was recorded. As it was a self-assessment, this data was not considered to be highly reliable. There are also ethical questions surrounding the comparison of educational background with age and performance on a memory game. As such no tests were performed using the educational background data. Computer experience may also have been a confounding factor, although most of the participants considered themselves to have average computer experience or better (indeed, most of the older participants were recruited via e-mail). It is possible that there are differences between the older and younger groups that confounded the experiment, but there is no clear evidence of any affect from computer experience on the results.

4.8 Guidelines

This study provided important information about the process of delivering notifications in different modalities, including important information about the effects if low-value (distractor) notifications. This section presents the findings refined into simple guidelines that will help developers who aim to use these modalities.

- The modality of an interruption does not affect how disruptive that interruption is. However, older people may be more susceptible to cross-modal interference, so care should be taken to avoid sensory channels being used by ongoing tasks.

- Visual and audio modalities should be given preference for most notifications, unless there are special circumstances (*e.g.* sensory impairment).

- For simple messages, abstraction appears to have little effect (once training has taken place). The training used in this study (described in Section 4.5) was not intensive, yet the abstract visual and audio modalities (*i.e.* abstract-visual, earcons) performed as well as the explicit modalities (*i.e.* text, speech). Abstract modalities can be used to deliver private information in a more discreet manner.

- Tactile notifications have some type of lag which manifests as a slower delivery time, a longer processing time, and a difficulty in differentiating distractor notifications from target notifications. While the real-world effects are not large, this should be considered when designing tactile notifications.

- Olfactory notifications should generally be avoided, as they produced the longest delivery times and the poorest response accuracy scores. However, performance was not so poor as to prevent olfactory notifications from being used in certain scenarios (*e.g.* to deal with severe impairment, or for non-urgent messages). Given their poor performance, olfactory notifications should not be used for important messages.

- Distractor notifications will cause negative effects that are similar to useful information, and as such their presence should be minimised as much as possible.

- Age has a significant effect on the ability to respond to notifications while carrying out an ongoing task. If notifications are being developed for older participants, the negative effects of both interruptions and distractions will be more severe. This will be more pronounced if the tactile and olfactory modalities are used.

4.9 Conclusions

Thesis Question 2 aimed to evaluate how notifications delivered in different modalities would affect users and activities. This study has addressed Thesis Question 2 by providing useful information about the properties of notifications in different modalities, specifically how disruptive, distracting and effective they are to ongoing tasks. The study also highlighted the effects of age; older participants reported higher subjective workloads, matched fewer cards and correctly responded to fewer notifications. While several effects were observed as a result of age, the results showed the performance patterns observed for older participants were similar to those observed for younger participants, *e.g.* both groups showed the 'tactile lag' effect. Both groups had trouble with tactile and olfactory notifications, although response accuracy levels of 70-80% show that these modalities are still capable information delivery mechanisms. The subjective workload scores revealed that older participants, as with younger participants, did not find any particular modality to have a significantly greater workload. Modality was also not found to effect primary task performance, again in line with younger participants. Reflecting on Thesis Question 1, the findings suggest that tactile and olfactory notifications should not be used to deliver important or urgent home care reminders unless there are special circumstances, *e.g.* significant visual and audio impairment. However, this work shows that should those special circumstances arise, tactile and olfactory methods are capable (at least with the simple messages used in this study).

The results suggest that there are 6 modalities (text, pictograms, abstract-visual, speech, earcons and auditory icons) which are capable of being used as primary interaction methods and 2 modalities (tactile and olfactory) that would be suitable for secondary interaction methods. Differences within the visual and audio methods, such as salience and abstraction/explicitness, can be used to deliver notifications in a more intelligent way by selecting the modality 'on the fly' to suit various situations and message types (e.g. delivering a sensitive reminder in an abstract method). The performance data gathered by this study will be vital to making informed design decisions when addressing Thesis Question 3.

In conclusion, the study presented in this chapter has addressed Thesis Question 2 by revealing important performance data on a range of modalities. This information can be used to identify which modality should be used given a different set of requirements, *e.g.* if the notification should be highly salient, then audio notifications should be used. Several researchers have proposed technology that is able to automatically switch between different modalities [8, 137, 171]. This results of this study will be used in Chapter 5 to guide the design of a dynamic multimodal system that explores how Thesis Question 3 could be answered.

Chapter 5

Dyna-Cue: A Multimodal Reminder Prototype

Thesis Question 3 was *"how can home reminder technology be designed to best utilise multiple types of interaction?"*. Several researchers have argued for technology that is able to automatically switch between modalities based on the environment and user's activities [9, 137, 171, 202]. Chapter 2 showed that there was a large amount of research to guide the design and applications of different modalities, yet there was not enough research that covered their relative performance properties to allow for the creation of a system that could make intelligent decisions about which one should be used in a given situation. This research gap was addressed by Chapters 3 and 4, which addressed Thesis Question 1 and 2 to provide data about the properties of different modalities, along with guidelines about where and when they should be used. To answer Thesis Question 3, this chapter presents the design and implementation of a prototype dynamic multimodal reminder system called Dyna-Cue, which aims to demonstrate that dynamic multimodal reminders can improve the quality of home care technology.

The requirements for Dyna-Cue were based around existing work and the recommendations made by other researchers; this is discussed in Section 5.1. An overview of the system that describes how the requirements were met is given in Section 5.2. The Dyna-Cue system has three primary subsystems: the Modality Registry (MR), discussed in Section 5.3; the Context Interface (CI), discussed in Section 5.4; and the Reminder Scheduler (RS), discussed in Section 5.5. Finally, Section 5.6 presents a summary of and conclusion to the work presented in this chapter. Appendix E presents an overview of the accompanying materials, which includes the full source code for the Dyna-Cue system.

This chapter considers the Dyna-Cue system in the home context using five activities and three short tasks. The activities are: sorting socks into matching pairs, watching television, listening to the radio, shopping from a catalogue and sorting household expenses. The short tasks are to cook food, eat food and take medicine. This context was used in the experiment presented in Chapter 6, and as such it is the configuration of Dyna-Cue used in that experiment which is used to demonstrate the model in this chapter. The selection of these activities and tasks is discussed in detail in Section 6.2.2 of Chapter 6.

5.1 Identifying the Requirements

Chapter 2 discussed several of the issues surrounding home care technology. Section 2.1 considered the requirements for home care technology and outlined several important factors, suggesting that it should be:

Reliable and Robust: Home care technology should be reliable [57, 207]. This is vital, as it is charged with helping to maintain the health and well-being of a person. In the case of failure, it must degrade gracefully.

Customisable: Home care technology must consider the unique needs of the user, which might include cognitive, mobility or sensory impairments [62, 142, 207].

Flexible: Home care technology must be flexible enough to handle developing conditions and impairments, changes to the environment and changes to home routines [62, 137, 165, 207]. This is generally called evolution. Edwards & Grinter [57] suggested that *all* technology for the home should be designed with evolution in mind.

Acceptable: Technology should be acceptable to the inhabitants of the home to prevent rejection [137, 142]. Two of the most prominent theories in the literature for making acceptable home technology are taking personal preferences into account [138, 143, 207] and to support/augment/utilize existing domestic routines [50, 48, 57, 193].

Accessible: Technology should be programmable or customisable without the aid of an engineer or special training [57].

Section 2.4 focussed on reminder systems and concluded that the best way to satisfy these requirements was to make technology multimodal and dynamic. The multimodal aspect of this is the concept of using more than one sensory system to communicate. With multiple modalities available for communication, the system has several different ways to interact with the user; this would help overcome sensory impairments, as non-impaired channels can be used for communication. Similarly, if delivery in one channel fails (*e.g.* due to hardware problems or saturation in that sensory channel) then another channel can be used, making the system more robust and allowing for graceful degradation. This could also help to address acceptability issues by using preferred modalities. Multiple modalities will also assist in evolution; if the requirements change over time, then alternative interaction methods are already a part of the system. Many researchers have advocated for technology equipped with multiple modalities for these reasons [9, 143, 162, 171, 226, 228]. It has also been shown that multiple modalities are superior to single modalities when interacting with impaired and older users [59, 108].

Interruption management is a concept that involves using information about the user and the environment to improve the effectiveness and acceptability of notifications. Most existing work is focussed around *when* to interrupt [11, 94, 152]. This work has demonstrated the value of prioritising and delaying notifications until a suitable 'interruption point' appears, such as when the user is between tasks. Less common is work on *how* to deliver a notification, but work has been carried out that shows acceptability can be improved by managing salience [140, 202] and politeness [10].

The Dyna-Cue system was designed to be dynamic in that it could use context information about the user, the task and the environment to make intelligent scheduling decisions. Dyna-Cue was also equipped with multiple modalities to help address many of the requirements set out above. One notable exception was configuration of the system; for the prototype, making an accessible configuration system was considered to be outside the scope of the work. However, the Dyna-Cue system was designed to be extensible to allow for this in future iterations. The primary requirements for the Dyna-Cue system were as follows:

R1: Multiple modalities must be available for delivering messages. Dyna-Cue must be able to automatically combine individual modalities to create multimodal notifications.

R2: The delivery methods must be independent to the logical parts of the system, so that modalities can be added and removed as needed without additional work.

R3: Dyna-Cue should react accordingly to changes in the environment or the user's activities that could impair the effectiveness of its delivery methods.

R4: Dyna-Cue should monitor the environment for compliance with notifications.

R5: If a notification does not produce the desired effect, the notification should be redelivered. Redelivery should use an alternative modality to the original message, in case the original modality was the point of failure.

R6: Dyna-Cue must prioritise the most important messages for delivery.

R7: The Dyna-Cue system should balance effectiveness and acceptability when sending reminders. This should be done by considering the environment, user activity and importance of the message when deciding *how* and *when* to deliver a reminder.

The following section describes how Dyna-Cue was designed to satisfy these requirements.

5.2 Overview of the Dyna-Cue Prototype

The requirements outlined in the previous section can be split into three groups. There are requirements relating to the management of modalities (R1, R2), requirements relating to the observation of the environment and user (R3, R4), and requirements relating to the decision-making logic in the system (R5, R6, R7). The Dyna-Cue prototype consists of three main components, each of which addressed one group of requirements.

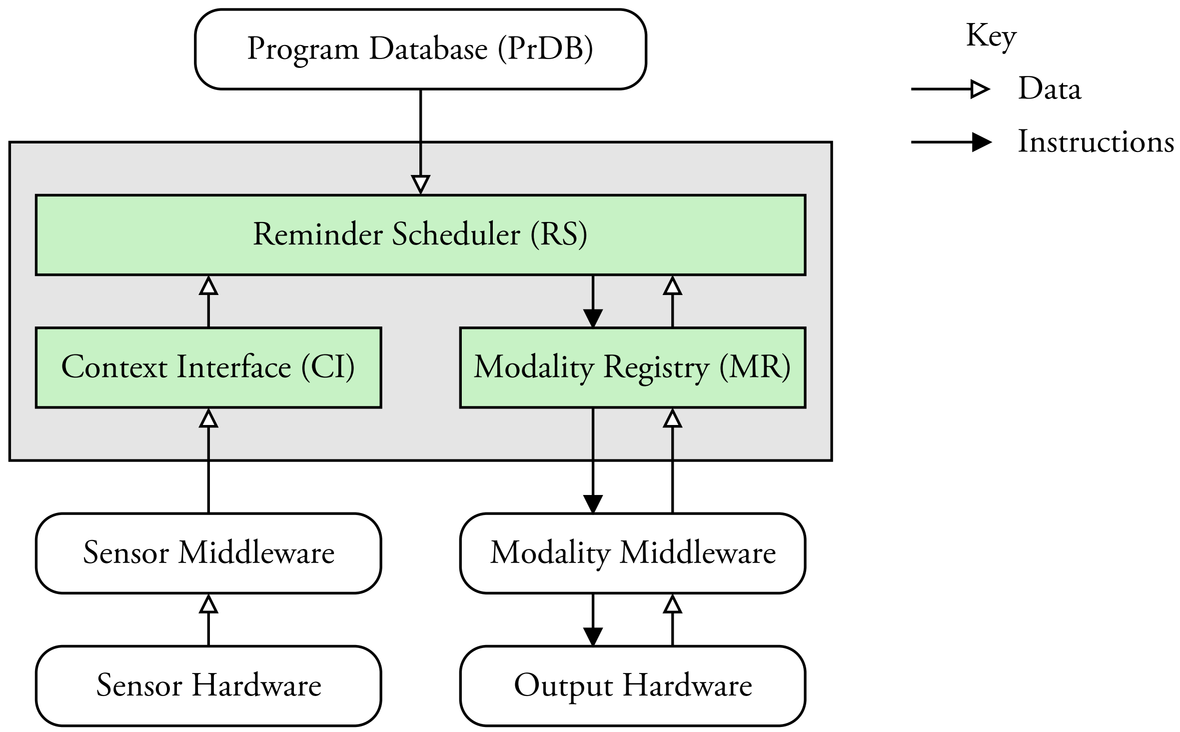


Figure 5.1: An overview of the Dyna-Cue prototype. The grey box represents the core of the Dyna-Cue system, with the items outside representing the inputs and outputs of the system. This diagram also shows the flow of control and data throughout the system. All the internal components are self-contained; instructions and data are passed between components, but none have direct control over the others. The Program Database (PrDB) is essentially a database that defines what the Reminder Scheduler (RS) should do in a given situation.

The first part is the Modality Registry (MR), described in Section 5.3, which is responsible for managing the modalities in the system. The second component is the Context Interface (CI), discussed in Section 5.4, which monitors the environment and user. The third and most important component is the Reminder Scheduler (RS), which manages the other two components and makes scheduling decisions. The Reminder Scheduler is presented in Section 5.5.

An overview of the design of the Dyna-Cue prototype is shown in Figure 5.1, which outlines how the components interact with each other. Figure 5.1 also highlights that the behaviour of the Dyna-Cue system is dictated by components that exist outside the core system. There are 3 external 'systems' that Dyna-Cue communicates with to function:

Modalities: There is a modality middleware layer that deals with the output hardware and provides specific functionality to the MR. This is discussed in more detail in Section 5.3.

Sensors: Sensors provide data about the environment and the user, and communicate with the CI. This is discussed in more detail in Section 5.4.

Program Database (PrDB): The PrDB is a database of events that Dyna-Cue should respond to, along with the message to deliver and some measure of how important that event is. The prototype can recognise several different types of events, as discussed in Section 5.5.

When Dyna-Cue starts, two configuration options take place. Firstly, the RS loads the behavioural program from the PrDB, which contains instructions on how the system should behave. The MR loads the individual modalities, and after confirming that the hardware is available, compiles a list of delivery methods available to the system. At this point the Dyna-Cue system starts the process of watching for notification events to respond to, which is described in more detail in Section 5.5.

The Dyna-Cue prototype was written in Java 1.7. Full source code is available in the accompanying materials, described in Appendix E.

5.3 Modality Registry (MR)

The purpose of the Modality Registry (MR) is to maintain a list of delivery methods and their properties. To clarify the terminology that will be used for the remainder of this chapter; a *modality* is the name given to a particular type of interaction (*e.g.* speech, text). A *delivery method* uses one or more modalities to delivery a given message to the user. Delivery methods can be unimodal or multimodal.

Figure 5.1 shows that the MR sits between the Reminder Scheduler (RS) and the modality middleware, which exists outside of the system. The modality middleware layer is actually a series of programs, one for each of the modalities the MR has access to. Each of the middleware programs serves two functions: *(1)* to manage the output hardware used by that modality, and *(2)* to provide data about that modality's properties to the MR.

The MR has several obligations, as follows:

- To manage the adding/removal of modalities to/from the system;
- Maintain a list of all possible delivery methods;
- Automatically combine modalities to create new delivery methods when possible;
- Report to the RS the properties of each of the delivery methods;
- Evaluate whether a given delivery method is capable of delivering a given message;
- To initiate message delivery using a given delivery method when instructed to do so by the RS.

There are two issues to address to provide this functionality. The first is the translation of messages from the RS to the specific output delivered by the different modalities; this is addressed in Section 5.3.1. The second issue is the scoring of the modalities and automatically combining them into delivery methods; this is discussed in Section 5.3.2.

5.3.1 Mode-Independent Representation

Reitter *et al.* [174] suggested that technology with multiple modalities could store messages in a 'mode-independent representation' then translate them at delivery time into a format suitable for the given modality. This provides flexibility as modalities can be added and removed when needed without additional work, assuming that each modality also includes some message-to-modality translator. The MR system was built around this design with each modality acting as a plug-in.

A notable advantage of this is that the modalities themselves are completely abstracted from the rest of the system, even within the MR. From the point of view of the RS, the delivery methods are represented solely by their numerical performance ratings. It also forces the modalities (*i.e.* the middleware programs representing the output hardware) to conform with a specific generalised interface, so the MR itself would not need to be reprogrammed to add new modalities to the system.

However, there is a wide range of information bandwidth between different modalities. A textual or synthetic speech modality could theoretically present any message, while an abstract modality such as earcons or tactons can only deliver messages that the user has been trained to understand. One of the functions of the MR is to evaluate whether a given modality is capable of delivering a given message.

For each delivery method the MR also maintains a list of messages it is able to deliver. In effect, this is the message-to-modality translator; messages are represented by keywords, and the modality's configuration file tells it how to deal with that specific keyword. The implementation of this is unique to each modality. All modalities capable of delivering a message are interchangeable, although they will have different performance properties.

When the RS asks the MR to deliver a message, the MR pares down its internal list of delivery methods to those capable of delivering that message. The resulting list is then returned to the RS, which evaluates their delivery properties to decide which method should be used. Once the RS has selected a delivery method it instructs the MR to initiate delivery of the message. This process is discussed in Section 5.5.

5.3.2 Modality Management

Automatically combining the modalities to create multimodal notifications required that all the modalities were entirely abstracted down to a series of comparable factors. To accomplish this, a series of challenges had to be overcome as follows:

1. Finding a way to abstract the modalities and represent them as simple numerical properties;

2. Creating a set of modalities that have been abstracted in this way;

3. Defining a set of rules that govern how the the scores of delivery methods are calculated from their component modalities.

This section details how these challenges were addressed.

Abstracting Modalities

Abstracting the modalities required defining a set of numerical values of interest when making scheduling decisions (*i.e.* figuring out the information required to make a decision on which modality should be used to deliver a message). Existing work has already identified several important factors in deciding which modality to use, as discussed in Section 2.4.

Latorella [123] found that audio notifications appear to be more disruptive to audio tasks and called this cross-modal interference. While it is not clear how significant this is for other modalities, Hoggan *et al.* [91] found that background noise would decrease the effectiveness of audio notifications, and similarly background vibrations for tactile. Hoggan *et al.* concluded that multimodal technology should attempt to use sensory channels with lower levels of environmental interference. Therefore, modalities must declare which sensory channel they use for interactions. This will also be important when combining modalities; while multimodal interactions have advantages over single modalities [59, 108], multiple notifications in the same modality delivered at the same time are likely to interfere with each other.

Perhaps the most important measurement that each modality should provide is some measure of how effective it is at delivering messages. For the purpose of the Dyna-Cue prototype, this was taken as a combination of how quickly the modality delivered the message and how easy it was to correctly interpret the message. While it may be useful to provide a range of effectiveness data (*e.g.* interpretation accuracy, delivery time and processing time were some of the measurements used in the previous chapters), for the purpose of the prototype a generalised value was desired.

Salience is a measure of how attention-grabbing something is. Vastenburg *et al.* [202] showed that more salient methods should be used to deliver more important information, and less salience methods for less important information. Therefore, the modalities in the Dyna-Cue prototype declared a salience value.

McGee-Lennon *et al.* [143] noted that sensitive information could be delivered by more abstract modalities to help prevent embarrassment to the user in social situations. Highly abstract methods also require training, and while the work in Chapters 3 and 4 showed that participants had a good success rate with abstract modalities, explicit notifications would be more appropriate if the message was highly important. Therefore the modalities in the Dyna-Cue prototype also declared an abstraction value that represented how explicit the message delivery was (or alternatively, how easy it was to intercept or eavesdrop on the message).

One of the primary research questions evaluated in Chapters 3 and 4 considered the relationship between modality and disruption. The results of those studies showed that the modality of the interruption had no effect on the disruption caused. Tactile interaction was a possible exception to that, however the effects appeared to only manifest temporally, *i.e.* the time spent trying to understand the notification. The real-world effect was very small. Other work has only been able to show a relationship between modality and disruption in terms of cross-modal interference [91, 123]. Modalities declare which sensory channel they use, and as such there is no need for modalities to also declare some 'disruptiveness' value. The sensory channel is sufficient to avoid cross-modal interference.

Based on this, when each modality registered with the MR it provided the following pieces of information:

- The sensory channel used (visual, auditory, *etc.*);
- An effectiveness rating, representing the ability of the modality to deliver information;
- A salience rating, representing how attention-grabbing the modality is;
- An abstraction rating, representing how much the modality obscures its meaning.

Configuring Modalities

Eight modalities were evaluated in Chapters 3 and 4. Due to the practical issues surrounding olfactory notifications it was not included in the Dyna-Cue prototype. The abstract-visual and earcon modalities were also not included in the Dyna-Cue prototype to avoid the need to train participants when evaluating the system (see Chapter 6). Although tactile notifications were included in the prototype, they were simplified to avoid the need for training.

Five modalities were implemented in the Dyna-Cue prototype: text, pictograms, speech, auditory icons and simple tactile vibrations. This section describes how these were configured in the Dyna-Cue system using the findings of the previous experiments. The configuration presented here is based on the assumption that the user has no significant sensory impairments. In the prototype MR each modality was given an abstraction, salience and effectiveness rating between 1 and 3 (with 1 being low and 3 being high). The modalities also declared whether they used visual, audio or tactile channels; this was represented as a binary value.

The tactile notifications used in the Dyna-Cue prototype were simple vibrations, unlike the notifications used in Chapters 3 and 4 which were structured tactile messages (tactons). This change was made for three reasons: *(1)* training was not desired, and would not be needed for a simple vibration as there was no 'message' being delivered; *(2)* the work in Chapters 3 and 4 concluded that most interactions should be use the

Table 5.1: Table showing a summary of the performance properties of the 5 modalities, based on the study presented in Chapter 4.

		Disruption			Effectiveness		
		Perf	R_a	R_e	T_d	T_p	*Acc*
Younger	Text	-9.90%	3.24%	34.01%	1.90	1.67	95.00%
	Pictogram	-9.48%	0.84%	26.69%	1.60	1.57	97.15%
	Speech	-6.88%	0.15%	18.24%	1.83	1.69	97.55%
	Aud. Icon	-6.77%	3.38%	26.08%	1.69	1.50	98.80%
	Tactile	-13.54%	-0.97%	22.09%	2.62	2.26	89.95%
Older	Text	-13.68%	1.39%	29.62%	2.26	1.85	90.63%
	Pictogram	-9.29%	-0.51%	18.00%	1.78	1.91	91.13%
	Speech	-9.77%	-3.70%	5.91%	1.89	1.95	95.88%
	Aud. Icon	-10.55%	1.45%	23.31%	2.06	1.98	95.94%
	Tactile	-13.24%	-2.11%	11.05%	3.02	2.82	79.88%

Note: *Perf*, R_a and R_e show the percentage difference between that modality and performance in the control condition (which had no notifications). *Perf* is overall performance (lower is better), R_a is activity rate and R_e is error rate. T_d is delivery time in seconds. T_p is processing time in seconds. *Acc* is response accuracy (percentage of notifications correctly acknowledged). The younger group was aged 18-30 and the older group was aged 50 and over. These measurements are discussed in more detail in Section 4.2.5.

Table 5.2: Table showing the base configurations of the modalities included in the Dyna-Cue prototype.

Modality	Abstraction	Salience	Effectiveness	Visual	Audio	Tactile
Text	1	1	2	1	0	0
Pictogram	2	1	2	1	0	0
Speech	1	2	2	0	1	0
Auditory Icon	2	2	2	0	1	0
Tactile	3	2	1	0	0	1

Note: The Abstraction, Salience and Effectiveness measures are rated 1 (low) to 3 (high). The Visual, Audio and Tactile values are binary and represent the sensory channel used by that modality. This table shows *all* the data provided to the system by each modality, demonstrating that the modalities themselves were completely abstracted.

visual and audio modalities for unimpaired users, with tactile serving as a support modality (*e.g.* by making notifications more salient); and *(3)* both Manly *et al.* [133] and McGee-Lennon *et al.* [146] have shown that simple abstract messages such as beeps can still serve as effective reminders (although not as effective as more explicit reminders).

Effectiveness measurements were based entirely on the results of the first two experiments, where visual and audio modalities were shown to be the most effective in terms of delivery speed and response accuracy as shown in Table 5.1. However, other work has shown that multimodal notifications are more effective than unimodal notifications [59, 108]. To compensate for this, the unimodal visual and audio modalities were rated at 2 (medium effectiveness) and the tactile 3 (low effectiveness) to allow multimodal notifications to be rated at 1. The ratings of the unimodal delivery methods are shown in Table 5.2.

Abstraction values were based on the explicitness of the modalities. Text and Speech are both highly explicit, so were graded 1 (low abstraction). Pictograms and Auditory Icons are both more abstract, but are not completely abstract; therefore they were graded as 2. Tactile messages in this case are completely abstract, as there is only one signal for all the messages with no way to understand its meaning. Therefore, the tactile message was rated as 3 (highly abstract). This is shown in Table 5.2.

The performance data from the earlier studies suggested that all the modalities were highly salient with the exception of Olfactory. Tactile demonstrated a lower response rate (see Section 3.7.2) than the visual and audio modalities in Chapter 3. However, the results from the second study in Chapter 4 showed significant and consistent levels of disruption for the tactile condition along with longer processing times. This suggests that in the cases where no response was provided, the participant failed to respond to the notification in time, not that they were unaware of the notification. Therefore, tactile was initially assumed to be as Salient as the other modalities; all were initially rated at 2 (medium salience). Similarly to effectiveness, multimodal notifications are more salient than unimodal notifications [120]. As such the high salience score was 'reserved' for multimodal delivery methods. One other adjustment was then made to the salience scores; the visual methods were downgraded to 1 (low salience). This decision was made based on two observations: *(1)* that attentional focus was required for the visual notifications, *i.e.* that if users were looking in the wrong direction they would not see the notifications; and *(2)* that research has shown that when carrying out a primarily auditory task visual notifications are given a lower priority [123, 164]. The final values for salience are shown in Table 5.2.

Table 5.2 shows all the data that the individual modalities provided to the MR, demonstrating that the delivery methods were entirely abstracted from the rest of the system. The Dyna-Cue prototype is evaluated in Chapter 6, which includes a reflection on the validity of these values in Section 6.7.

Creating Delivery Methods

The MR used the data shown in Table 5.2 to automatically create unimodal and multimodal delivery methods by aggregating the scores of the individual modalities. The rules for combining the modalities was defined by existing work, some of which was covered in the previous section. The following rules were used to determine which modalities to combine and to calculate their ratings:

R.1: Modalities of the same sensory channel should not be combined to prevent cross-modal interference ([91, 123]).

R.2: Combinations of modalities are more effective than individual modalities; however, there is little return on using more than two modalities ([59, 108]).

R.3: Multimodal methods are more salient than their unimodal equivalents ([120]); *e.g.* text and speech together will be more salient than text or speech individually.

R.4: A multimodal method is only as abstract as its least-abstract modality (*e.g.* a text-tactile reminder is not abstract, although it contains an abstract modality).

These rules were expressed as code in the Dyna-Cue prototype as shown in Listing 5.1. Taking the individual modalities shown in Table 5.2 as input, these rules produced the delivery methods and scores shown in Table 5.3. Using these rules the 5 individual modalities were combined to create 5 unimodal, 8 bimodal and 4 trimodal delivery methods. With 5 modalities the MR was able to provide 17 delivery methods to the RS. Based on these rules, new modalities can be added and removed from the system with minimal work.

Listing 5.1: Code sample showing how modalities were combined in the system to produce several notification modalities.

```
//Combine the modalities within a delivery method
for(Modality a:modalities){
    abstraction     = Math.min(abstraction,
                               a.getStat("Abstraction"));
    salience        = Math.max(salience,
                               a.getStat("Salience"));
    effectiveness   = Math.max(effectiveness,
                               a.getStat("Effectiveness"));
    visual          = Math.max(visual,
                               a.getStat("Visual"));
    audio           = Math.max(audio,
                               a.getStat("Audio"));
    tactile         = Math.max(tactile,
                               a.getStat("Tactile"));
}

//Increase stats if the delivery methods use more than one modality
if(modalities.size()>1) salience++;
if(modalities.size()>1) effectiveness++;
```

Table 5.3: Table showing the full scores for the delivery methods used in the Dyna-Cue prototype. A, S and E show the abstraction, salience and effectiveness scores respectively. Vis, Aud and Tac are binary values showing whether that delivery method uses the visual, auditory or tactile sensory channels.

Delivery Method	A	S	E	Vis	Aud	Tac
Text	1	1	2	1	0	0
Pictogram	2	1	2	1	0	0
Speech	1	2	2	0	1	0
Aud.Icon	2	2	2	0	1	0
Tactile	3	2	1	0	0	1
Text-Speech	1	3	3	1	1	0
Text-Aud.Icon	1	3	3	1	1	0
Text-Tactile	1	3	3	1	0	1
Pictogram-Speech	1	3	3	1	1	0
Pictogram-Aud.Icon	2	3	3	1	1	0
Pictogram-Tactile	2	3	3	1	0	1
Speech-Tactile	1	3	3	0	1	1
Aud.Icon-Tactile	2	3	3	0	1	1
Text-Speech-Tactile	1	3	3	1	1	1
Text-Aud.Icon-Tactile	1	3	3	1	1	1
Pictogram-Speech-Tactile	1	3	3	1	1	1
Pictogram-Aud.Icon-Tactile	2	3	3	1	1	1

5.4 Context Interface (CI)

The purpose of the Context Interface (CI) is to provide information on the current activity of the user and the state of the environment. This information would normally be interpreted from sensor data, such as the ones used the Gator Tech Smart House [85, 121, 125] or the House_n [18, 17, 99, 100, 157] projects. However, a context-aware sensor network lies outside the scope of this work. Instead, a simple interface was developed that allowed the manual input of context data in real time.

The CI provides five pieces of information to the RS, as follows:

- A keyword for the current activity;
- The time spent on the current activity;
- Current saturation of the user's visual channel;
- Current saturation of the user's aural channel;
- Current saturation of the user's tactile channel.

For example, if the user is watching TV then the CI would report that the user is watching TV, the length of time they have done so and that the task is primarily visual and audio. In the Dyna-Cue prototype, the RS was also able to request that the CI alert it when single events took place (*e.g.* taking medicine, turning on the oven, *etc.*). In those cases the CI would transmit the name of the event and the time that it occurred to the RS. This allowed the RS to monitor several different types of behaviour. The advantages of this are discussed in Section 5.5.

In the Dyna-Cue prototype the implementation of the CI was quite basic. A set of activities were preconfigured and loaded into a simple interface as shown in Figure 5.2. When the 'TV' task was selected, the data for this activity was loaded and transmitted to the RS. For the prototype this was done using a configuration file, making it easy to change the behaviour of the system. In a real-world system this could be replaced with a context-aware sensor network. While it is possible that the user could self-report their activities, which would require very little alteration of the existing system, self-reporting is unlikely to be suitable for use in a home-care context.

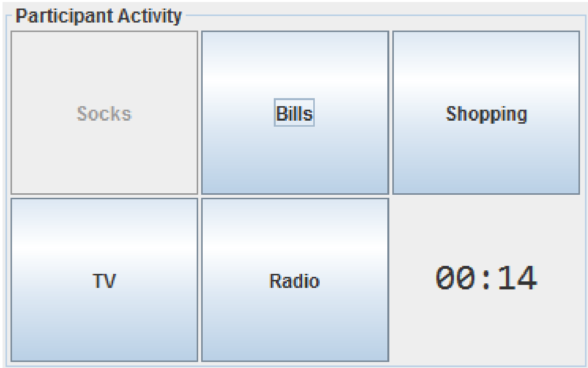


Figure 5.2: The activity selector that served as a substitute for a sensor network in the Dyna-Cue Context Interface prototype.

5.5 Reminder Scheduler (RS)

The part of the prototype that controls the delivery of reminders is the Reminder Scheduler (RS). The RS is the core of the Dyna-Cue system and decides how and when to deliver notifications. As shown in Figure 5.1, the RS interfaces with three systems: the Context Interface (CI), which provides information about the user and their activities; the Modality Registry (MR), which provides information about the delivery methods available for use; and finally the Program Database (PrDB), which is an external database containing instructions for the RS that determine its behaviour.

The first step that the RS takes is to load the operational data from the PrDB, which is discussed in Section 5.5.1. After this is completed the RS will regularly query the CI for user activity updates. When an event occurs that requires a notification, the RS takes note of it and works out the desired properties of the message's delivery based on how important that message is. This process is detailed in Section 5.5.2. The RS will then query the MR to form a list of delivery methods capable of delivering the message. The RS will then score the delivery methods based on their performance properties, discussed in Section 5.5.3. The RS will then attempt to deliver the message and will monitor the CI for signs of compliance. The message will remain in the system until some signs of compliance have been detected; if they are not, then the RS will attempt to redeliver the message. This process is discussed in Section 5.5.4. An overview of the RS's behavioural processes are shown in Figure 5.3.

5.5.1 The Program Database (PrDB) & Schedule

The Program Database (PrDB) is a database that contains a set of events to the scheduler should react to be sending a reminder notification. When the Dyna-Cue

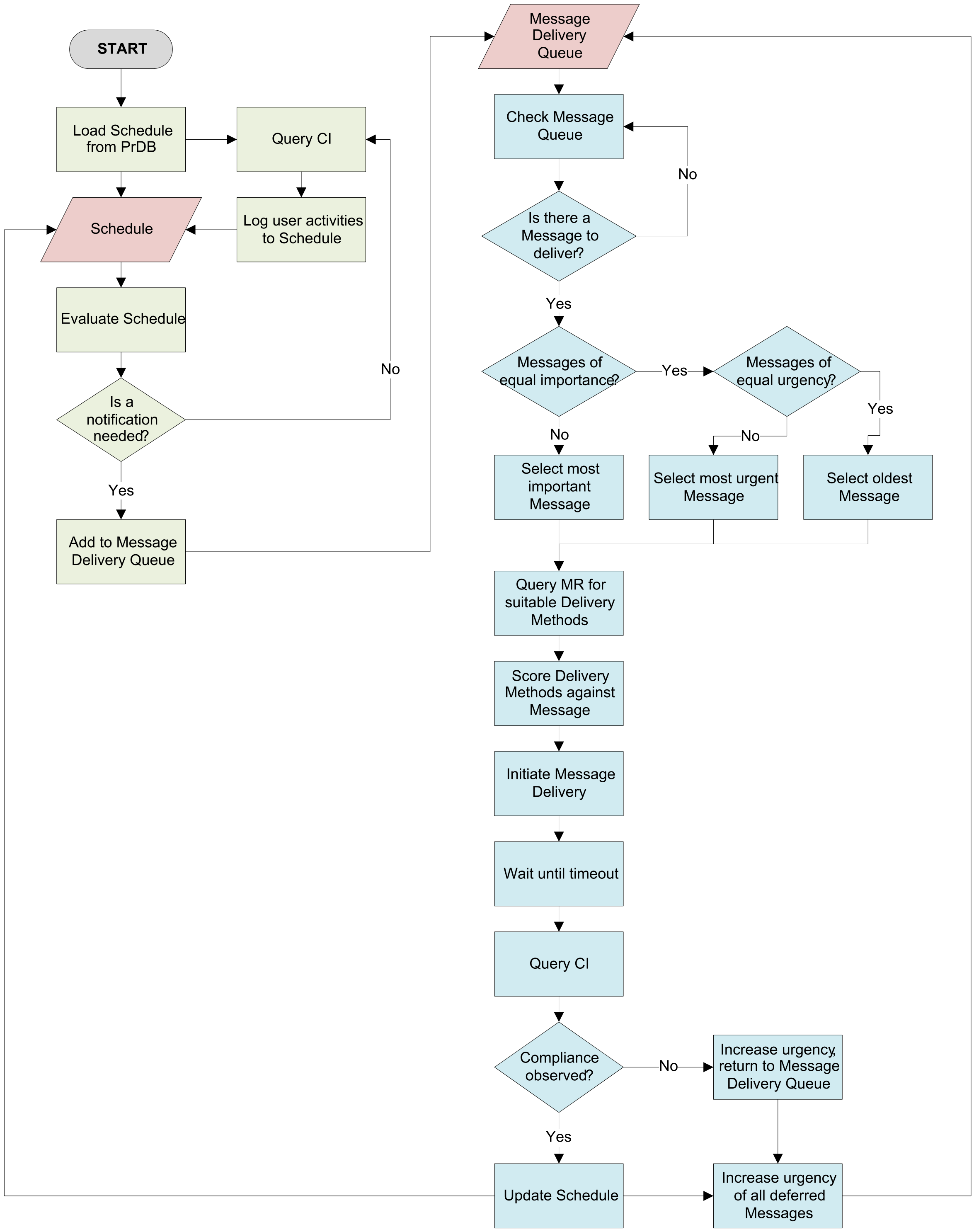


Figure 5.3: Flowchart showing how the Dyna-Cue RS handles message delivery. The red boxes represent data. There are two parallel processes shown. The green boxes show the process of tracking the user and deciding whether or not to send a reminder. The blue boxes show the process of deciding which message to deliver, when to deliver it and how.

system is started, the database is loaded in to the *Schedule*. The Schedule is an 'in memory' database that contains two sets of data: *(1)* the event data loaded from the PrDB and *(2)* a record of the user's behaviour. The RS constantly checks the CI and adds data to the Schedule. Every time the Schedule is updated, the RS checks to see if the observed behaviour deviates from the desired behaviour. If so, it starts the process of delivering a reminder. This is process is shown in Figure 5.3.

The PrDB and Schedule are able to identify and respond to a range of complex behaviours. There are two types of event; *ongoing* events and *immediate* events. Ongoing events have a start and end, while immediate events simply have a completion time (*i.e.* they are either completed or not completed, they cannot be 'partly complete'. Some examples of ongoing events are watching TV, doing laundry, jogging, *etc.* Examples of immediate events are taking medicine, locking doors, turning on the oven, eating, *etc.*

Surrounding these events are different types of trigger, which the RS uses to identify deviations from intended behaviour. The triggers included in the Dyna-Cue prototype are as follows:

Time Started: If set, the time that an event is to be started at. Can only be used for ongoing events.

Time Completed: The time that an item is to be completed by. This applies to both ongoing and immediate events.

Duration: The time that should be spent on an activity. Applies only to ongoing events.

Relative To: If set, then the related start/complete times are set relative to the start/complete times of another event. This can be applied to both ongoing and immediate events.

This can be used to create quite complex chains of trigger conditions, such as:

At 5:00pm, watch TV for an hour. At 6:00pm, eat. One hour after eating, take medicine.

If an event is triggered, then a Message is created. A Message in the system contains a mode-independent representation of the message to deliver and an importance value. This data are provided by the PrDB.

5.5.2 Classifying Messages

When the RS decides that a notification is needed, it creates a Message as defined in the previous section. The message is created with a mode-independent representation of the information to deliver and a value representing the importance of that information. Like the measurements used in the MR (see Section 5.3), the importance value ranges from 1 (low importance) to 3 (high importance).

Messages that are deferred or unacknowledged are likely to be more urgent than new messages entering the system with the same importance score. This could be addressed by increasing the importance score of a message the longer it exists in the system. However, this could cause problems as a low-important message that has been waiting for some time could be perceived as equally urgent to a new high-important message. However, the high-important message should take priority. For example, a reminder to water plants should never be considered as important as a fire alarm.

To resolve this, a second measurement is attached to the message called urgency. The urgency score is used to calculate all other properties of the message. Therefore, the initial urgency score is set to be equal to the importance score of the message, ensuring that the most appropriate methods are used for delivery.

To clarify, the *importance* score is used to decide which message takes priority (*i.e.* when it should be delivered). The importance score does not change. The *urgency* score is used to determine how a message should be delivered (*i.e.* which delivery method to use). The urgency score increases the longer the message has been in the system, and its initial value is equal to the importance score.

When a message is chosen to be delivered, the desired properties of the delivery method are worked out. This involves two steps: *(1)* checking the CI to find out about the current environment and activity, then *(2)* considering the desired levels of abstraction, salience and effectiveness for the delivery method. These values are determined based on the following rules:

Table 5.4: Table showing the desired abstraction, salience and effectiveness values for three different levels of urgency.

Urgency	Abstraction	Salience	Effectiveness
1	3	1	1
2	2	2	2
3	1	3	3

Listing 5.2: Code that defines notification parameters.

```
this.desiredAbstraction      = 4-urgency;
this.desiredSalience         = urgency;
this.desiredEffectiveness    = urgency;

this.currentVisualWorkload   = currentVisual;
this.currentAudioWorkload    = currentAudio;
this.currentTactileWorkload  = currentTactile;
```

- The more important the message, the less abstract the message should be to ensure the user understands it. With a less important message, more abstract modalities should be used as they are still effective (based on the work presented in Chapters 3 and 4) but can preserve privacy [143].

- The more important the message, the more salient it should be; less important messages should also be less salient [202].

- The more important the message, the more effective the delivery method should be to ensure that the message is delivered.

These rules were expressed in the prototype as shown in Table 5.4. Salience and effectiveness increase with urgency while abstraction decreases. That salience and urgency are the same suggests that only one of these measures is necessary. However, these values are treated differently when scoring takes place (which is discussed in Section 5.5.3). They also have different meanings, despite their scores, and keeping both measurements helps to ensure that future work will have both sets of information available. Listing 5.2 shows how the message parameters were calculated within the Dyna-Cue prototype.

Each message now has 6 pieces of data attached to it: the desired levels of abstraction, salience and urgency for the delivery method and the current visual, audio and tactile engagement of the user. This data are used to score the delivery methods as shown in the following Section.

5.5.3 Scoring Delivery Methods

To deliver a reminder Dyna-Cue first has to select the appropriate delivery method from the options provided by the MR. Firstly, the RS would ask the MR to provide a list of the delivery methods that could deliver the message. The RS then takes the resulting list and scores them based on their abstraction, salience, effectiveness and modality requirements. The rules used for this calculation were as follows:

- The closer to the desired salience, the better; the score should be lower the further the delivery method's salience is from the message's requirements. The same applies to the abstraction scores. In other words, a message should not be more or less abstract/salient than it needs to be.

- The delivery method chosen should be *at least* as effective as the message requires; all other things being equal, the scheduler should pick the most effective delivery method for the best performance.

- Notifications should avoid using sensory channels that are currently engaged or have significant background interference ([91, 123]).

All of the delivery methods were scored individually against the requirements of the message and the current saturation/engagement of the 3 sensory systems (visual, audio and tactile). The score for each modality was calculated as follows:

- The score starts at 0.

- For every property which matches the message requirements, the score increases by 1.

- If abstraction or salience does not match, the absolute difference between the desired and actual values is subtracted from the score.

- If the desired effectiveness is lower than the actual effectiveness, the absolute difference is subtracted from the score.

These rules were coded into the Dyna-Cue prototype as shown in Listing 5.3. Based on the 17 delivery methods created for the prototype (see Section 5.3) and some common household activities (defined as part of the evaluation in Chapter 6), Figure 5.4 shows

Listing 5.3: Code that calculates the score for a delivery method to send a particular message.

```
public int calculateScore(Notification n){
    // Abstraction
    int abs = (abstraction == n.desiredAbstraction) ?
        1 : -(Math.abs(abstraction - n.desiredAbstraction));
    // Salience
    int sal = (salience == n.desiredSalience) ?
        1 : -(Math.abs(salience - n.desiredSalience));
    // Effectiveness
    int eff = (effectiveness >= n.desiredEffectiveness) ?
        1 : -(Math.abs(effectiveness - n.desiredEffectiveness));
    // Visual
    int vis = (visual != n.currentVisualWorkload) ?
        1 : 0;
    // Audio
    int aud = (audio != n.currentAudioWorkload) ?
        1 : 0;
    // Tactile
    int tac = (tactile != n.currentTactileWorkload) ?
        1 : 0;
    // Sum parts and return total score
    return (abs + sal + eff + vis + aud + tac);
}
```

this process in its entirety. When the delivery methods have been scored, the highest scoring method is selected for delivery, and a delivery request is sent to the MR.

5.5.4 Delivering Notifications

When the highest scoring delivery method has been chosen, the RS requests that the MR attempts to deliver the message. The MR will then watch for signs of compliance from the CI. In the Dyna-Cue prototype, this was straightforward as the CI sends data about the current task, and also sends alerts when it registers that an immediate event has occurred. If the RS detects compliance, then it logs this and marks that event as being completed.

If no signs of compliance are observed, then Dyna-Cue attempts redelivery. The message's urgency score is increased and the MR is queried for delivery methods again. With the increased urgency scores, it is unlikely that the same modality will be chosen again for re-delivery, which is illustrated by Figure 5.4. This process continues, with urgency increased at every re-delivery request, until compliance is detected. If compliance is not detected then the notification will remain in the system and the RS

Task	Visual	Audio	Tactile
TV	1	1	0
Bills	1	0	1
Radio	0	1	0
Shopping	1	0	1
Socks	1	0	1

Urgency	Abstraction	Salience	Effectiveness
1	3	1	1
2	2	2	2
3	1	3	3

Delivery Method	Text	Pictogram	Speech	Aud. Icon	Tactile	Text+Speech	Text+Aud. Icon	Text+Tactile	Pictogram+Speech	Pictogram+Aud. Icon	Pictogram+Tactile	Speech+Tactile	Aud. Icon+Tactile	Text+Speech+Tactile	Text+Aud. Icon+Tactile	Pictogram+Speech+Tactile	Pictogram+Aud. Icon+Tactile
Abstraction	1	2	1	2	3	1	1	1	1	2	2	1	2	1	1	1	2
Salience	1	1	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3
Effectiveness	2	2	2	2	1	3	3	3	3	3	3	3	3	3	3	3	3
Visual	1	1	0	0	0	1	1	1	1	1	1	0	0	1	1	1	1
Audio	0	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1
Tactile	0	0	0	0	1	0	0	1	0	0	1	1	1	1	1	1	1

Activity	Urgency	Desired Abstraction	Desired Salience	Desired Effectiveness	Currently Visual	Currently Audio	Currently Tactile	Text	Pictogram	Speech	Aud. Icon	Tactile	Text+Speech	Text+Aud. Icon	Text+Tactile	Pictogram+Speech	Pictogram+Aud. Icon	Pictogram+Tactile	Speech+Tactile	Aud. Icon+Tactile	Text+Speech+Tactile	Text+Aud. Icon+Tactile	Pictogram+Speech+Tactile	Pictogram+Aud. Icon+Tactile
TV	1	3	1	1	1	1	0	1	2	-1	0	4	-3	-3	-1	-3	-2	0	-1	0	-2	-2	-2	-1
Bills	1	3	1	1	1	0	1	1	2	1	2	2	-1	-1	-3	-1	0	-2	-1	0	-2	-2	-2	-1
Radio	1	3	1	1	0	1	0	2	3	-2	-1	3	-2	-2	0	-2	-1	1	-2	-1	-1	-1	-1	0
Shopping	1	3	1	1	1	0	1	1	2	1	2	2	-1	-1	-3	-1	0	-2	-1	0	-2	-2	-2	-1
Socks	1	3	1	1	1	0	1	1	2	1	2	2	-1	-1	-3	-1	0	-2	-1	0	-2	-2	-2	-1
TV	2	2	2	2	1	1	0	0	2	2	4	2	-1	-1	1	-1	1	3	1	3	0	0	0	2
Bills	2	2	2	2	1	0	1	0	2	4	6	0	1	1	-1	1	3	1	1	3	0	0	0	2
Radio	2	2	2	2	0	1	0	1	3	1	3	1	0	0	2	0	2	4	0	2	1	1	1	3
Shopping	2	2	2	2	1	0	1	0	2	4	6	0	1	1	-1	1	3	1	1	3	0	0	0	2
Socks	2	2	2	2	1	0	1	0	2	4	6	0	1	1	-1	1	3	1	1	3	0	0	0	2
TV	3	1	3	3	1	1	0	-1	-3	0	-2	-2	3	3	5	3	1	3	5	3	4	4	4	2
Bills	3	1	3	3	1	0	1	-1	-3	2	0	-4	5	5	3	5	3	1	5	3	4	4	4	2
Radio	3	1	3	3	0	1	0	0	-2	-1	-3	-3	4	4	6	4	2	4	4	2	5	5	5	3
Shopping	3	1	3	3	1	0	1	-1	-3	2	0	-4	5	5	3	5	3	1	5	3	4	4	4	2
Socks	3	1	3	3	1	0	1	-1	-3	2	0	-4	5	5	3	5	3	1	5	3	4	4	4	2

Figure 5.4: Diagram showing how Dyna-Cue combines basic information to make scheduling decisions. This diagram shows several parts of the system. The top-left table shows the basic abstraction, salience and effectiveness values desired for different levels of message importance (Section 5.5.2). The table to the bottom-left shows how this is combined with some test environmental data from common household tasks (discussed further in Chapter 6). The top-right table shows how the individual modalities were combined to make bi-modal and tri-modal delivery methods (Section 5.3.2). Finally, the bottom-right table shows how the different delivery methods were scored based on the parameters from the bottom-left table. A higher score is more likely to be selected by the Dyna-Cue scheduler while a lower score is less likely. This is highlighted by the colour-coding of the table.

will continually attempt to redeliver the notification, randomly switching between the top-scoring delivery methods. This process is shown in Figure 5.3.

Conflicts

It is possible for RS to be asked to deliver multiple messages at the same time. In this situation, the RS first examines which message has the highest importance rating. The most important message is given priority, and a buffer around message delivery ensures that only one message is delivered at a time. That is, less important messages are deferred, in the same manner as Iqbal & Bailey [104], Horvitz *et al.* [94] and Vastenburg *et al.* [203, 202]. Note that deferred messages will increase their urgency scores at the same rate as messages which are not deferred, reflecting the time they have been waiting in the system.

If messages have the same importance values (which do not change), then the urgency values are used instead (which increase the longer the message has been in the system). Therefore, messages which have been in the system the longest will be prioritised. If two messages of equal importance enter the system at roughly the same time, then the message which enters the system first will take priority. Both messages will increase in urgency at the same rate, but the second message (despite having the same scores as the first) will be blocked until the first event is resolved. The implications of this are discussed in Section 6.7.

5.6 Conclusions

Thesis Question 3 considered how a home reminder system could utilise multiple interaction methods. The Dyna-Cue prototype presented here represents one possible solution to this problem. Dyna-Cue attempts to make intelligent scheduling decisions for more appropriate and effective reminding. Section 5.1 argued that the over-arching requirements for home care technology as reliability, customisation, flexibility, acceptability and accessibility. Dyna-Cue satisfies the requirements of robustness, customisation and flexibility: the multiple modalities provide several options for customisation and flexibility, which are automatically deployed if one of the methods fails to work correctly. The system's behaviour can be changed by modifying the schedule stored in the PrDB.

Sensory impairment can be taken into account by adjusting the scores for each of the modalities (*e.g.* by reducing the effectiveness scores of impaired sensory channels).

The prototype system does not satisfy the requirement of accessibility, as changing the behaviour of the system currently requires a programmer or engineer. While this is an acceptable shortcoming for a prototype, further research would be needed to fully understand how to enable end-user configuration in a fully-implemented system. End-user configuration of home care technology is an area of active research [98, 137, 208] that lies outside the scope of this work.

The requirement of acceptability is difficult to evaluate, as the Dyna-Cue prototype would need to be evaluated before this could be known. The Dyna-Cue system does attempt to make interactions more appropriate by following the findings of other researchers, *e.g.* balancing salience with importance as suggested by Vastenburg *et al.* [202]. However, it is not clear if the system as a whole could be considered more acceptable; it would need to be assessed against similar systems before any conclusions could be drawn.

In conclusion, the Dyna-Cue prototype addresses Thesis Question 3 and satisfies most of the requirements for home care technology set out in Section 5.1. However, the Dyna-Cue prototype must be evaluated to fully understand if this sort of technology is more effective and appropriate than existing models of reminder delivery. This is addressed in the following chapter.

Chapter 6

Evaluating the Dyna-Cue Prototype

Thesis Question 4 was *"can home reminder technology be made more effective and appropriate by providing it with the ability to dynamically select from multiple forms of interaction?"*. Chapter 5 described the creation of a home reminder system prototype called Dyna-Cue. Dyna-Cue was designed to address some of the problems faced by home care technology, such as the difficulty in compensating for sensory impairment and requirements that change over time. The overall aim of the Dyna-Cue system is to balance the effectiveness (how well it works) with acceptability (a positive evaluation from its users).

This chapter addresses Thesis Question 4 by assessing the Dyna-Cue prototype against other types of reminder technology and the unassisted human brain. Section 6.1 sets out the aims of the study. The design of the study is described in Section 6.2, followed by Section 6.3 which specifies the hypotheses that were tested by the study. The participants used in the study are described in Section 6.4 followed by the procedure in Section 6.5. The results are given in Section 6.6 and discussed in Section 6.7. The primary findings of this study are refined into guidelines in Section 6.7 and this chapter concludes with Section 6.10.

6.1 Aims

The aim of the study was to answer Thesis Question 4 by evaluating the Dyna-Cue prototype described in Chapter 5. As discussed in Section 5.1, the reason the Dyna-Cue prototype was created was to address some of the shortcomings of existing reminder technology while demonstrating the value of dynamic technology. More specifically, the prototype was created to demonstrate that a dynamic multimodal scheduler could be *(1)* as effective as other types of reminder technology while *(2)* being more appropriate for use in a home care setting.

As the aim of a reminder system is to help users manage their daily lives and activities, the evaluation needs to test the ability of the Dyna-Cue system against similar reminder systems. Perry *et al.* [171] discussed dynamic multimodal technology and suggested that there could be many problems with such systems, such as the high potential for confusion when switching modes (modalities). The evaluation will explore these issues by gathering subjective feedback and comparing the Dyna-Cue prototype to a similar system which does not switch modalities. Therefore, the evaluation will compare the Dyna-Cue prototype to alternative reminder delivery mechanisms in order to answer the following research questions:

Research Question 1

How does the Dyna-Cue system compare to other reminder delivery methods with respect to helping people to manage time and organise activities?

Research Question 2

Does the Dyna-Cue prototype produce more appropriate (*i.e.* correct and acceptable in the user's opinion) interactions compared to alternative reminder delivery methods?

Research Question 3

Does the Dyna-Cue prototype create a higher workload compared to alternative reminder delivery methods?

The work in Chapter 4 revealed that age had a significant effect on the performance of the older group, and therefore it is important to include older users in the evaluation. It is also important to include younger participants, as reminder systems are not exclusively for the use of older people.

Answering these research questions would provide vital information to guide the development of multimodal reminder technology. If dynamic multimodal reminders can be shown to help address issues often found in home care scenarios, such as high levels of background interference or sensory impairment, then home care technology could be made more robust, effective and acceptable to its users.

6.2 Design

Based on the aims of the evaluation, testing required a mixed-models design with age as a between-subjects variable and different types of reminder technology (one of which would be the Dyna-Cue system) as repeated measures variables. The study required a task that provided ecological validity while also providing a reliable measurement that can be used to compare the abilities of the different technologies.

A review of existing tests is presented in Section 6.2.1, followed by Section 6.2.2 which describes the task used in the experiment itself. Section 6.2.3 describes the different types of reminder that were delivered to participants in some of the conditions. Section 6.2.4 specifies the independent variables, Section 6.2.5 the dependent variables, and Section 6.2.6 discusses the confounding variables and their potential to impact on the experiment.

6.2.1 Prospective Memory Tests

Home reminder technology should help home care recipients to manage their daily activities and tasks. Therefore, any experiment aiming to evaluate home reminder technology should be based organising/remembering tasks and activities. While it would be logical to base the experiment around something like *Activities for Daily Living* (ADLs) [219], which are used to decide how much care a person needs, ADLs are activities such as bathing and eating which are difficult to replicate in an experiment in a meaningful way.

Reminder systems came into existence primarily to counteract a specific symptom of natural ageing: cognitive decline [41]. While all forms of memory decline naturally with age [169], prospective memory is particularly prone to decline and has a significant effect on daily living [188]. Smith *et al.* called prospective memory *“memory for* *intentions"* [188, p. 311]; memory that is used to remember when and what to do in certain situations and at certain times.

There are two types of prospective memory: event-based and time-based. Event-based (also called *externally-cued*) prospective memory describes remembering to do something after a prompt or clue, such as remembering to post a letter when you see a post office. Time-based (also called *internally-cued*) prospective memory describes remembering to do something at a particular time, such as remembering to turn off a tap when running a bath, or taking medicine at the correct time. Groot *et al.* stated that *"failures in prospective memory . . . can have devastating effects on everyday life"*, and that *"forgetting to do things could threaten independent living"* [74, p. 645].

Frontal-lobe brain damage also causes prospective memory decline (although generally at different levels of severity) [74, 188]. As a result, there are several clinically validated tests that can be used to assess task organisation and timekeeping. One of these tests would be ideal for the evaluation of the Dyna-Cue prototype.

There is also another factor to consider, however: ecological validity. Burgess *et al.* [38] argued that ecological validity is vital for the development of better clinical tests, and that the results from ecologically valid studies are generally more reliable than the results of clinically validated but less ecologically valid studies. Therefore, to maximise the reliability of the evaluation a test should be selected that represents life at home. This is of particular importance given that one of the aims is to gather subjective data.

The Wisconsin Card Sorting Task

Created by Berg [20], this is a well-known Prospective Memory test that asks participants to sort cards in a particular way. While sorting cards, the rules are changed and the participant is expected to adapt to sorting the cards in the new way. The participant is not notified of any mistakes, and the number of mistakes made is the primary measure of this technique. This can be quickly administered, is simple to explain, and is well-known[1]. However, the task lacks ecological validity as it is not home-based.

[1] Google Scholar lists 1,006 citations for the original paper.

The Six Elements Test

The Six Elements Test was first presented by Shallice and Burgess [183]. Participants attempt to complete 6 tasks split into 3 categories: dictating a route, arithmetic and writing. Tasks were of similar difficulty. Participants have 15 minutes to attempt to carry out the tasks but are also given a set of rules that dictate when and how they can carry them out. The participants are assessed on the number of tasks they carry out and the maximum time spent on each task. Participants could view a clock at any time to help them organise their time.

The Six Elements Test is much like the Wisconsin Card Sorting Task, in that it is portable and quick to administer. The primary problem with the test is that it lacks a connection to the home, and it is not clear how it could be modified to provide one. The constraints placed on the tasks are also somewhat artificial.

There is also a revised version called the Modified Six Elements Test, created by Wilson *et al.* [222]. This test is still based around pathfinding, arithmetic and writing but is designed for people with more severe frontal lobe damage and is far less demanding. Subjects are expected to attempt each task within a 10 minute time limit, with the rule that they cannot work on two similar tasks consecutively.

The Modified Six Elements Test was designed as part of a well-known test battery called the Behavioral Assessment of Dysexecutive Syndrome (BADS) [222]. BADS also includes a variant of the Wisconsin Card Sorting Test, questionnaires, and a test called the Zoo Test. The Zoo Test is a paper-based trip planner and logic puzzle. The BADS test battery is interesting but takes a long time to administer, is difficult to alter for a home context, and is designed for people with significant frontal lobe damage (*i.e.* people with symptoms far more severe than would normally be found in age-related cognitive decline).

The Multiple Errands Test

The Multiple Errands Test was created by Shallice & Burgess [183]. Participants are given 8 tasks to accomplish in a pedestrian shopping area; the first 6 are simple, *e.g.* buy some bread. The 7th task asks them to be in a certain place at a certain time, while the final task is more challenging, tasking them with finding out information (*e.g.* which shop sells the most expensive item). Participants should spend as little money as possible and complete the tasks as quickly as they can. As in the Six

Element Test, participants have rules they need to follow about how the tasks should be completed. The result of this test included checking for specific types of error, such as misunderstanding the rules.

This test gathers useful information ‘in the wild’, and also provides an insight into the role of social interactions in dealing with prospective memory decline. However, the test does not have a home context, the test is complex, and there are several confounding variables that could alter the outcome as a result of carrying out the test in public.

The Hotel Test

The Hotel Test was created by Manly *et al.* [133] to demonstrate how auditory cues could help people with brain injuries organize their activities. Based on the Modified Six Elements Test, participants had 15 minutes to attempt 5 tasks related to the running of a hotel: sorting coins, proof-reading, sorting name tags, looking up phone numbers and compiling bills. In addition to this, participants had to open and close garage doors (by pressing a button) at set times. All of these tasks were carried out at the same table in full view of a clock. The clock was covered so that the experimenter could take a note of when participants wanted to check the time.

Participants did not have enough time to complete each of the tasks; participants were instead to attempt to spend as much time as possible on each one (with a 15 minute limit, ideally 3 minutes for each task). Scoring was based around the number of tasks attempted, task time allocation and the correct and timely opening/closing of the garage doors. The aim of the Hotel Test is to evaluate timekeeping and organisation, so the performance on the different tasks was not important; however, this was not clear to the participants.

The Hotel Test is much more appropriate than the other tests discussed. It can be administered quickly, gathers data on task organisation, is portable, and has already been used to evaluate the effects of notifications on timekeeping. The Hotel Test has also been tested with healthy adults which shows it has an appropriate level of cognitive demand. The primary shortcomings of the Hotel Test are that it is not home based and that there is a learning effect (the Six Elements and Multiple Errands both used random tasks to reduce the impact of a learning effect). Both problems can be addressed in theory by replacing the hotel-style tasks with home-style tasks that can be randomly configured.

The Bungalow/Removal Task

Developed by Sweeney *et al.* [192], the Bungalow Task uses a virtual environment and asks the participant to play to the role of a furniture removal person. In a virtual home the participant has various prospective memory tasks to complete, including *(1)* closing doors behind them, *(2)* checking outside for the removal van's arrival every 5 minutes and *(3)* labelling glass items with fragile stickers. Furniture has to be tagged for removal to a new house, but has to be tagged in a particular order; lounge furniture must be tagged first, *etc.* Measures included rule-breaks, time-based, event-based and activity-based prospective memory lapses, strategy and total rooms visited. There is no time limit.

This is the first of the tests which has a home-based context, but it is far removed from regular home-style activities. While the test itself gathers a lot of useful and insightful data, the lack of a time limit makes for poor experimental controls. The virtual environment could be used to provide an alternative home-based context, but would be a poor substitute for *real* home-based tasks that could be carried out in a lab setting.

Discussion

The evaluation will be carried out in a lab setting to allow controls, but must provide some level of ecological validity. This can be done by asking people to carry out the same sort of tasks they would normally carry out in their own home. The most appropriate prospective memory test would be the Hotel Test; it can be administered quickly, it measures the ability to organise time and intentions, and has previously been used to test the effects of alerts. In fact, for the Modified Six Elements Test (of which the Hotel Test is a variant) the authors state:

> “ *An unusual aspect of this test is that it is not important how well the subject performs in the individual components. For example, it does not matter how accurate the picture naming is or how many of the arithmetic problems the subject manages to do. The point of the test is to measure how well subjects organise themselves.* ”
>
> — Wilson *et al.* [222, p. 219]

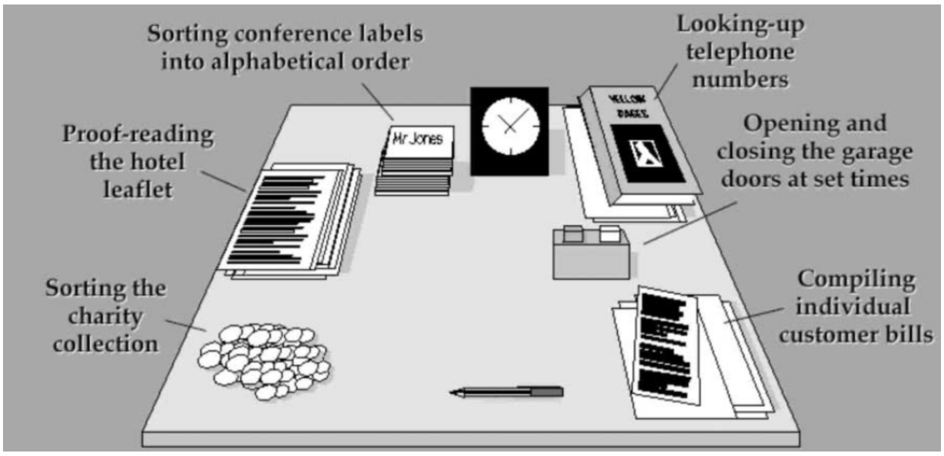


Figure 6.1: The Hotel Test as described by Manly *et al.* [133].

So could the experimental tasks be replaced with ones with a home-based context without harming the validity of the design? Both Wilson *et al.* [222] and Manly *et al.* [133] set out the same requirements for the experimental tasks: they must be representative of complex everyday activities, and it should *not* be possible to complete them in the time allocated during the experiment. Several other researchers have argued that ecological validity is the most important factor in tests like the Six Elements Test, and that the experimental tasks should reflect familiar activities [38, 187, 192]. Groot *et al.* [74] even called the experimental tasks "filler tasks" due to the fact that performance is not measured.

Therefore, the most appropriate prospective memory test is a variant of the Modified Six Elements Test [222] or Hotel Test [133] given a home-style context. Due to the similarity between the aims of this study and the aims of the experiment by Manly *et al.* [133], the experimental design should primarily be based on the Hotel Test.

6.2.2 The Experimental Test

In the previous section the Hotel Test by Manly *et al.* [133] was shown to be an appropriate experimental test for the analysis. The Hotel Test is given the context of a job interview for a hotel, which is not an appropriate context for the evaluation. However, due to the design of the Hotel Test and the Six-Elements Test on which it is based, the hotel-style activities can be substituted for alternatives [183, 222]. The requirements for the experimental activities are:

1. that they should represent ecologically valid everyday activities;

2. that they cannot be completed in the allocated time period.

The original Six-Elements Test and the Hotel Test both provide partial external cues for prospective memory, as the tasks to be carried out are all visible. It is clear which tasks have been completed and which have still to be carried out. The Hotel Test added an additional layer of complexity on top of the original Six-Elements test by introducing a parallel time-based (or internally cued) activity, which was pressing a button at set times during the experiment. No cues were given to support this task; the clock was covered unless the participant actively looked at it, and there were no visual indicators to let participants know that they have completed the test.

The 5 tasks which are partly event-based (externally cued) were called the primary tasks. The parallel time-based (internally cued) task was called the secondary task. To make the Hotel Test suitable for the home, both the primary and secondary tasks should be replaced with home-style variants.

Primary Experimental Tasks

The primary tasks used in the Hotel Test were sorting coins, proof-reading, sorting labels, looking up phone numbers and compiling bills as shown in Figure 6.1. For each of these a home-based analogue needs to be found which cannot be completed in the allocated time and resemble the activities commonly carried out at home. A third requirement was also added for this study: that the five tasks use a range of modalities, instead of being primarily visual as they were in the Hotel Test. Like the Hotel Test, all the tasks had to fit on the top of a table, which imposed some restrictions on the type of tasks that could be carried out.

As the focus is changing from a work environment to a home environment, it would be ecologically valid to include *passive tasks* such as watching TV instead of attempting to create a range of tasks with equal mental workloads. Given that this task will be used to evaluate a dynamic multimodal scheduler, the activities were also chosen based on the modalities they use. The 5 activities defined for the home task were sorting socks into matching pairs, listening to the radio, watching TV, home shopping with a catalogue and calculating household expenditure.

Laundry Task
The coin sorting task was a visual reasoning task, so it was replaced with a similar visual reasoning task often carried out at home: sorting socks. It was predicted that

this task would be very familiar to almost all participants regardless of age. A total of 36 pairs of socks were put into a box and mixed around. The socks were a mixture of various dark colours and stripes to make sure that this task was sufficiently challenging. The task was piloted with 3 people, none of whom were able to match all the socks in less than 3 minutes. The socks task provided a similar level of mental workload to sorting coins while being firmly grounded in the home context.

Shopping Task

The shopping task replaced the Hotel Test's phone book task, which involved looking up phone numbers in a telephone directory. While it is true that people at home do often look up phone numbers at home, it was not considered to be a regular household activity, and it is not a task that would normally take longer than 3 minutes. Initially it was replaced with an online shopping task, as both would be visual searching tasks. Due to the space that a computer would take up on the table along with a desire to keep the tasks non-technical, the computer aspect was replaced with a shopping catalogue[2]. A series of shopping scenarios and budgets were created, such as:

> *"A friend's child is having their 7th birthday soon. You want to get them and educational gift with a budget of £25."*

The full list of budget scenarios are provided in Appendix C. This task is very similar to the original Hotel Test activity but suitable re-purposed for the home.

Budget Task

The Hotel Tesk included a bill-compiling task, which was intended to be an arithmetic test. Calculating a household budget was the closest home-based analogue that matched the requirements of the original activity, an expenses and budgeting task was devised. This required a budget sheet along with a set of fake receipts and invoices. The budget sheet is included in Appendix C along with a description of how the receipts were created. There are too many receipts to include here as a new set was required for every condition: Figure 6.2 shows an example set of receipts. All receipts are included in the accompanying material (see Appendix E. To carry out the Budget Task participants were provided with a pencil, budget sheet and a collection of receipts. The budget sheet has categories such as 'groceries', 'bills', 'nights out', *etc.* and participant were asked to classify the expenses and then calculate total expenditure.

TV Task

With the Laundry, Shopping and Budget Tasks defined only two tasks remained from the original Hotel Test: sorting name badges and proof-reading, both tasks rarely carried

[2]Specifically, the Argos UK Spring/Summer 2012 catalogue.

(a) All receipts. **(b)** Close-up of receipts.

Figure 6.2: The receipts used in the household expenses activity. Figure a shows all the receipts that were used in a single trial, with a fresh set of receipts used in every condition. Figure b shows a closeup of the receipts.

out at home. Given the focus on finding tasks that used multiple sensory channels and tasks that were ecologically valid, it was decided that the remaining two tasks would not attempt to map directly onto their Hotel Test counterparts. Given the near ubiquity of televisions, it was a clear choice for an ecologically valid task that employed an alternative set of modalities. While the other tasks so far have been visual-tactile, watching TV is visual-audio.

The TV was controlled by the experimenter and used a computer monitor as shown in Figure 6.3a. The TV was also used to display visual notifications as shown in Figure 6.4; this is discussed in more detail in Section 6.2.3. As watching TV is a passive task, it might not produce the same Zeigarnik 'tendency to complete' effect [230] that is likely to influence how and when people chose to carry out the tasks. To combat this, an attempt was made to select 'interesting' clips from TV shows that people would enjoy, as follows:

- Stephen Fry's opening speech at the 2012 BAFTA awards;
- A clip from Top Gear UK in which Jeremy Clarkson repeatedly crashed a 3-wheeled Robin Reliant;
- A stand-up routine from 'Live at the Apollo' by comedian Jimeoin;
- A clip from the comedy "The IT Crowd" in which one of the characters is tricked into believing the Internet is contained in a small black box.

All the clips were at least 5 minutes long and were chosen to be entertaining, relatively neutral subject matter and without profane language. Television often serves as their main source of companionship for isolated elderly people, as highlighted by the Friends of the Elderly charity during "Isolation Week". This involved subjecting younger people to the levels of sensory deprivation and social isolation as are experienced by many elderly people.[3] As such, it was considered particularly important to include a TV-related task.

Radio Task

Given the sensory channel usage of the other tasks, an audio task was desired. Initially it was considered that people could be asked to write down a recipe played from a tape, however this idea was dropped as it would have meant all the tasks used the visual sensory channel. Instead, the Radio Task was reduced to simply listening to the radio. Participants were asked to select a radio station they liked at the start of the experiment[4] and simply asked to listen to it for three minutes. It was expected that while younger participants might not listen to the radio, they were very likely to listen to music around the home in some form and therefore the radio would be ecologically valid.

Secondary Experimental Tasks

One of the key points of the Hotel Test was the internally-cued prospective memory task, which was to open and close a garage door at appropriate times. When at home, people tend to have a large number of prospective tasks to perform such as remembering to turn off a tap, to take food out of the oven or to watch a TV show when it is broadcast. Such events are the primary target of reminder technology, and as such this portion of the test was expanded on for the home style variant.

Eating and taking medicine at the right time are often used as example applications for home reminder systems, and make for ideal secondary tasks in this context. Food is usually cooked before being eaten, and medicine is often taken before, with or after eating. This provides a straightforward set of secondary tasks: cook, eat and take medicine.

To reflect real life, the time that these events took place was not statically defined. The first task (cook) was defined to start at a certain time, while the remaining tasks (eat, take medicine) were defined relative to the previous task. To carry out the activities, the

[3] Friends of the Elderly, `http://www.isolationweek.com/home/`.

[4] Many participants had no preference, so Classic FM was used. This is `101.7 FM` in Glasgow, but varies country-wide. See `http://www.classicfm.com/` for details.

participant would press the appropriate button as shown in Figure 6.3b. Participants were given a card at the start of the experiment which defined the times that the buttons should be pressed, which they were allowed to keep at hand for reference during each condition. For example, the card might say the following:

- *After 4:30 minutes, you need to start cooking. Do this by pressing the 'cook' button.*

- *The food takes time to cook. Press the 'eat' button 7:00 minutes after you have pressed the 'cook' button.*

- *You need to take your medicine on a full stomach. Press the 'medicine' button 3:15 minutes after pressing the 'eat' button.*

The full set of instruction cards given to participants is provided in Appendix C.

Summary

The experimental task was based on the Hotel Test by Manly *et al.* [133] and asked participants to carry out a set of 5 primary tasks, dividing 15 minutes of time equally over each task (ideally spending 3 minutes on each task). As in the Hotel Test [133] (and all other variants of the 6 elements task [183, 222]) the tasks are designed to take longer than the allotted time to complete. There were 5 primary tasks, chosen to represent home-based activities carried out in various modalities. The tasks were:

1. Sorting clean socks into matching pairs;
2. Listening to the radio;
3. Watching television;
4. Shopping from a catalogue;
5. Calculating and categorising a household budget.

The tasks were arranged around the participant as shown in Figure 6.3a. The tasks themselves were not assessed; instead, the test measured a participant's ability to organise their time by examining how closely their activities matched a prescribed schedule. A clock was available to the participants to help them organise their time, which could be accessed at any time by pressing a button marked 'clock' (see Figure 6.3c).

Participant were also asked to carry out a secondary task, which was to press three buttons at set times during the 15 minute period. The buttons represented cooking, eating and taking medicine as shown in Figure 6.3b. These tasks were carried out relative to the preceding task, so participants had to compensate for being late or early.

6.2.3 Notification Design

As noted in Chapter 4, the Dyna-Cue system can be configured with a range of modalities. Five modalities were included for the purposes of this experiment: text, pictograms, speech, auditory icons and tactile vibrations.

Based on the results of the previous experiments (Chapters 3 and 4), it is clear that olfactory interactions should not be used for interaction under normal circumstances. The long delivery times associated with olfactory notifications rendered it entirely impractical for this experiment, as the window in which the participant was expected to respond was much shorter than the average delivery time for olfactory notifications. While there are several use-cases for olfactory notifications, they are not appropriate for this experimental design.

The most abstract audio and visual methods (earcons and abstract visual messages) were also excluded. However, this was not due to their performance: the work carried out in Chapters 3 and 4 showed that they were effective reminders. However, they would require significantly more training than the other visual and audio methods, for which minimal training was provided. As the final experiment required around 90 minutes, and as older participants were used in the experiment, the additional time cost of providing such training posed a significant practical problem. As the experiment would already include two visual and audio modalities with varying levels of salience, there was no clear benefit to offset this cost and as a result the most abstract visual and audio modalities were excluded.

Figure 5.4 shows how the Dyna-Cue system selected one of the 5 modalities based on the tasks outlined in Section 6.2.2. Each of the modalities had to be configured to transmit one of four different messages: cook, eat, take medicine and change task. This section defines how the modalities were configured for this purpose.

(a) Overview of Experimental Test.

(b) Secondary Task Buttons.

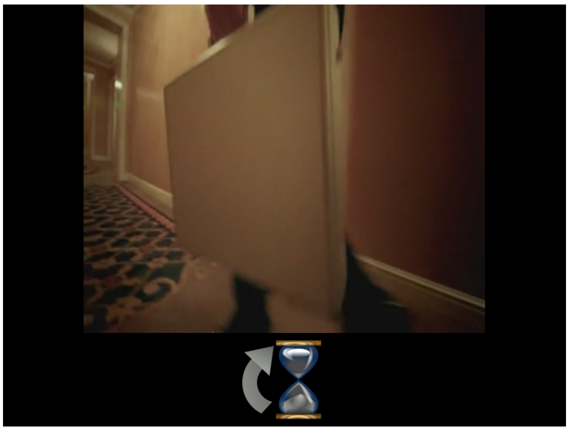
(c) Clock Button and TV Remote.

Figure 6.3: The full setup of the experimental test. Figure a shows, clockwise from the bottom-left: the Shopping Task, the Budget Task, the Laundry Task, the Radio Task and the TV Task. Figure b shows the 3 buttons used in the secondary task and Figure c shows the clock button. The mobile phone in the centre of Figure a is given to the participant to put in their pocket at the start of the experiment.

(a) Screen when watching TV.

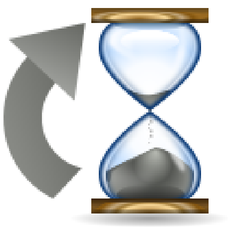
(b) Screen when pressing the clock button.

(c) Delivery of text notification.



(d) Delivery of pictogram notification.

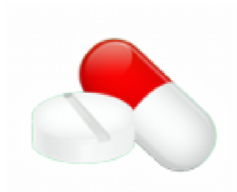
Figure 6.4: The TV and delivery of visual reminders. While these may appear small in this figure, they are much more salient when displayed on the 19” flat-panel display used in the experiment.



(a) Change Task

(b) Cook

(c) Eat



(d) Take Medicine

Figure 6.5: The pictograms used in the experiment.

Text

The TV, shown in Figure 6.3a, served multiple roles during the experiment. It would always be on and would either show a blank screen, a TV show (Figure 6.4a) or the clock (Figure 6.4b). It was also used to display textual messages however, as shown in Figure 6.4c. When a text message was delivered the current display was shrunk and the message was printed at the bottom of the screen in a white sans-serif font. The textual messages were as follows:

- It is time to change to another task.
- It is time to start cooking.
- It is time to eat.
- It is time to take your medicine.

Textual messages would remain on the screen for three seconds.

Pictograms

Four pictograms were used in the experiment as shown in Figure 6.5. These were displayed in the way as textual notifications, *i.e.* at the bottom of the TV display. This is shown in Figure 6.4d, which also shows that the pictograms take up much more vertical space than textual messages did (and therefore 'shrunk' the TV/clock more to compensate). This was to account for the larger horizontal spacing of the textual messages. If the pictograms had been shown at the same vertical size as the textual messages, then they would be very small and not particularly attention-grabbing. As the configuration of the Dyna-Cue system (see Figure 5.4) expects the textual and pictographic notifications to have an equivalent salience level, the size of the pictograms was increased to compensate. Like the textual messages, the pictograms stayed on-screen for 3 seconds.

Speech

Four speech messages are used in the experiment, all of which use the same synthetic speech engine from previous experiments (the 'Heather' voice from The Scottish Voice). The speech messages matched with the textual messages and were played from a laptop used by the experimenter.

Table 6.1: A description of the auditory icons used in the study.

Message	Auditory Icon
Change Task	A boxing bell ringing three times.
Cook	The sound of an onion being chopped quickly.
Eat	The sound of a person taking a bite out of a carrot.
Take Medicine	The sound of a pill being removed from a bottle by turning it on its side, then upright again.

Auditory Icons

Four auditory icons were used in the experiment as shown in Table 6.1. The auditory icons were sourced from an online sound effects repository and modified to suit the experiment with Audacity.

Tactile

Tactile interactions were the only fully abstract modality kept from those tested in earlier chapters. However, it was reduced in complexity to a single vibration pattern, acting as a very simple alert or as a ‘support‘ to more effective modalities (*i.e.* by attempting to increase their salience). The work presented in Chapters 3 and 4 found unusual properties for tactile notifications; participants appeared to forget them quickly after training, and there was an unusual lag in response times that could not easily be explained.

As the the length of the experiment resulted in the exclusion of abstract visual and audio modalities, it would also be advantageous if the training segment for tactile messages could also be eliminated. Previous work also showed that interpretation rates for tactile notifications were not very high, which is likely to be related to the difficulty in remembering the mapping of vibration patterns to actions.

In the original Hotel Test, Manly *et al.* [133] demonstrated that periodic beeps (not timed to the participant’s behaviour in any way) helped participants to manage their time by prompting them to consider their actions. As the effectiveness of such a notification has already been demonstrated, it suggests that a simple vibration might also be an effective reminder. When that reminder is delivered only when something is demanding attention, it is expected to be an effective reminder. In addition, with

Figure 6.6: The Nexus One phone used to deliver tactile notifications during the experiment.

no mapping between the vibrations and the expected response, no training would be required.

Chapters 3 and 4 also suggested that tactile vibrations could be used to increase the salience of other modalities; *i.e.* that the tactile vibration is simply to get attention, while the actual message is delivered through other means (much like a text message alert). Using a single vibration pattern instead of multiple patterns meant that a highly salient pattern could be chosen without worrying about how to define vibrations patterns that were distinctly different.

In conclusion, the simple tactile vibration is not intended to deliver a specific message, but rather to remind the recipient that there is something requiring their attention. It is able to deliver this payload immediately and effectively. As the message is only "*something needs your attention*", there is also little chance of an interpretation error, and the vibration could easily be combined with more explicit modalities.

Tactile notifications were delivered via an Android Nexus One Phone, as shown in Figure 6.6, connected to the experimental computer using Bluetooth. Participants were asked to put the phone into their pocket and remove their own phone (participants are always asked to turn off their phones). Participants were asked beforehand to wear trousers or jeans with participating in the experiment. The tactile vibrations fluctuated quickly from high to low intensity over a period of 1.5 seconds at the maximum intensity allowed by the device.

6.2.4 Independent Variables

There were two independent variables in the study: age and the type of assistance provided. Age was a between-groups variable while assistance provided was a repeated-measures variable.

Much like in Chapter 4, two age groups were desired for the experiment; a younger group and an older group. The younger participants group was aged 16-30 and the older participant group 60 and over. All participants were expected to be able-minded[5] and without significant sensory impairment. More information on the experimental participants is provided in Section 6.4.

There were four different types of assistance provided to participants, as follows:

No Reminders
No reminders were delivered to the participant. The participant has free access to the clock, which they are encouraged to use to manage their activities.

Static Reminders
The participant will receive reminders when they deviate from the schedule, however they will always be delivered using text and speech modalities.

Random-Modality Reminders
The participant will receive reminders when they deviate from the schedule, however the modality will be randomly selected.

Dynamic Reminders
The participant will receive reminders when they deviate from the schedule with the modality selected by the Dyna-Cue system as described in Chapter 5.

The static condition is intended to represent a current-generation reminder system limited purely to text and speech. Section 2.2 demonstrated that the current generation of home care technology is only starting to include speech technology; the majority of existing technology is still limited to visual interaction. The static 'text and speech' condition is intended to represent a home reminder system without *any* kind of multi-modal logic, using the same modalities in every situation regardless of the environment, social situation or user activity.

[5]That is, we did not actively seek out participants with cognitive impairments. Cognitive ability was not screened in any way.

The random condition was included to test for a multimodal effect. If the dynamic condition showed a significant difference to the static condition, would it be due to the effect of multimodality, or due to the dynamic logic? Including a random dynamic condition would help to answer that question.

Both the static and random conditions were created by reconfiguring the Dyna-Cue system. Therefore they both delivered the same messages, used the same schedules, and had access to the same data as the Dyna-Cue prototype. In practice, this means that the rules governing *when* to deliver a notification were identical for all three reminder conditions; it is only the rules governing *how* to deliver the information that changed between systems.

6.2.5 Dependent Variables

The participants in the Hotel Test were expected to spend exactly 3 minutes on each task. Every second under or over 3 minutes is counted and totalled over all tasks to give a value called the *task organisation score* (TOS). The lower the TOS, the better the participant organised their time over all the tasks. This was applied to both the primary and secondary tasks.

The primary dependent variable for the study was a more refined version of the TOS. This was calculated as in the Hotel Task but with some minor variations. Firstly, points that would have been impossible to avoid were not counted towards the total. For example, if the penultimate primary task is 20 seconds late, then the final task would have to be completed 20 seconds early. In this example, only 20 points were added to the TOS instead of 40. The second modification made to the scores was that events which occurred *ahead* of schedule were not counted, as the reminder system in its current form cannot prevent a person from doing something ahead of schedule; the prototype is only able to remind people about things they have failed to do on time. Therefore, only the *preventable* lapses were of interest in this evaluation.

Several subjective factors were of interest, as the second research question considered whether the Dyna-Cue system was more appropriate than the alternative systems. Participants were asked to rank the conditions in terms of how easy they were to use, the participant's preferences, which one was most appropriate for the home, and which one was most appropriate for the home. These were measured using simple ranking questions. Appendix C includes the questionnaire that was given to participants to measure this.

As with the previous studies in Chapters 3 and 4, the NASA Task-Load Index (NASA-TLX) was also of interest. As reported by Hart [78] the full test with weightings is not necessary for statistical validity, so for simplicity the one-page assessment survey without weightings was administered after each condition (Raw TLX).

To summarise, the dependent variables were as follows:

Task Organisation Score (TOS): a simple measure of how many seconds away from the ideal schedule the person was. A lower score signals better time management.

NASA-TLX: a measure of subjective workload.

Subjective Difficulty: measured by asking participants to rank the conditions from hardest to easiest.

Subjective Preference: measured by asking participants to rank the conditions in terms of their personal preference, from most favourite to least favourite.

Appropriateness for the Home: measured by asking participants to rank the conditions in terms of what they considered to be the most appropriate for use in the home.

Appropriateness for a Care Scenario: measured by asking participants to rank the conditions in terms of how appropriate it would be in a care setting.

The subjective assessments and NASA-TLX surveys were administered using paper surveys, details of which are provided in Appendix C. The TOS was measured automatically by the experimental computer, although data input to the system was carried partially in a 'Wizard of Oz' fashion. More details of how data was entered into the system are given in Section 6.5, which discusses the procedure of the experiment.

6.2.6 Confounding Variables

There were several possible confounding variables, but the most important variables were gender, participant apathy and familiarity to the tasks. Gender was considered an important factor to manage due to the popular opinion that females are better at multitasking than men [175]. To control for any possible gender effects, an even gender distribution was desired for both user groups.

Participant apathy was considered to be a more significant issue in this experiment than for similar experimental designs. This is because the participant is expected to believe that their performance on the tasks is important, and therefore they would be expected to attempt a high level of performance. In reality the study measured the ability to manage time between equally-important tasks. If participants knew that their performance on the tasks was not important, but their time-keeping was, it is likely that they would focus on their attention on timekeeping instead of the tasks themselves. If so it would result in behaviour that is unlikely to be ecologically valid. To help ensure that participants believed they were being assessed (despite the presence of two passive tasks), their results for the shopping and budget tasks were kept and the number of socks matched were counted and recorded in full view of the participants. As noted in Section 6.5, this appeared to be successful with most participants asking if they had improved on their previous performance.

Familiarity with the tasks was another potential confounding factor in two ways: *(1)* the tasks were chosen to be ecologically valid, so if the tasks are unfamiliar, then ecological validity has not been achieved; and *(2)* participants who were more familiar with the home-style activities might be expected to perform better on them. However, as performance in the activities was not a primary factor in the study, it is not clear what (if any) effect this would have on the results. The conformity of the home-style activities with regular household activities was measured by a subjective 5-point Likert scale, as shown in Appendix C. These data provided an important insight into the ecological validity of the experimental design and is discussed further in Section 6.7.

6.3 Hypotheses

This section outlines the hypotheses tested during the study. Note that as discussed in Section 6.2.4, the random condition was included in the experiment to provide an insight into any differences between the static and dynamic conditions. As such, none of the hypotheses make specific predictions about that condition, although it is still included in all the analyses presented in Section 6.6.

6.3.1 Research Question 1 – Task Organisation

Research question 1 asked "*how does the Dyna-Cue system compare to other reminder delivery methods with respect to helping people to manage time and organise activities?*".

This was tested by one variable: the task organisation score (TOS), as defined in Section 6.2.5.

Task Organisation Score – H 1.1

As discussed in Section 6.2.5, the TOS was measured by the number of seconds away from the perfect schedule (excluding 'impossible' seconds and 'early' seconds). It was predicted that the Dyna-Cue system would be at least as effective as the static system at delivering improving TOS. Both the dynamic and static conditions are expected to provide a significant improvement compared to the 'no assistance' condition. Age is expected to be a factor given natural age-related cognitive decline, but no interactions are expected. This gives the following hypotheses:

H 1.1.1 The dynamic and static conditions will both produce a lower TOS than the 'no assistance' condition, with the dynamic condition performing *at least as well as* the static system.

H 1.1.2 The age of the participant will affect the TOS.

H 1.1.3 There will be no interaction between modality and age on the TOS.

6.3.2 Research Question 2 – Appropriateness

Research question 2 asked "*does the Dyna-Cue prototype produce more appropriate (i.e. correct and acceptable in the user's opinion) interactions compared to alternative reminder delivery methods?*". This was testing by four subjective variables: difficulty, preference, appropriateness for the home, and appropriateness for care. These were measured by subjective rankings as defined in Section 6.2.5.

Subjective Difficulty – H 2.1

Difficulty was measured as a simply ranking. No subjective differences are expected between the static and dynamic conditions; both are expected to be voted easier than the condition without assistance. Age is not expected to be a factor, as the condition without notifications is expected to always be ranked 'most difficult' by both younger and older participants. The following hypotheses were tested:

H 2.1.1 The dynamic and static conditions will both be consistently ranked easier than the 'no assistance' condition, but will not be significantly different from each other.

H 2.1.2 The age of the participant will not affect the subjective difficulty rankings.

Subjective Preference – H 2.2

Subjective preference was also measured using a simple ranking as defined in Section 6.2.5. It was predicted that the dynamic condition would be the most preferred system for both younger and older people, as it has been designed with the aim of providing an appropriate type of interaction based on the context, while the other conditions have not. This was tested by the following hypotheses:

H 2.2.1 The dynamic condition will be preferred by participants.

H 2.2.2 Age will not have an impact on preference.

Appropriateness for Home – H 2.3

As with the other subjective measures here, appropriateness for the home was measured using a simple subjective ranking. It was predicted that the dynamic system, which favours subtlety at first, would be considered much more appropriate for the home than the static system. It may be that the younger participant group would not consider a reminder system appropriate at all, and might consider the 'no reminders' condition to be the most appropriate tool for the home. This is not clear so was tested with a two-tailed hypothesis. The hypotheses tested were as follows:

H 2.3.1 The dynamic system would be considered most appropriate for the home by older participants.

H 2.3.2 There will be a difference in the subjective assessment of appropriateness for the home between the two age groups.

Appropriateness for Care – H 2.4

Appropriateness for care was measured with a simple ranking. It was expected that the dynamic condition and the static condition would both be considered highly appropriate for a care scenario. The static condition was expected to perform well in this aspect as it is a somewhat 'heavy handed' way to deliver notifications, and it is expected that many participants would prioritise effectiveness over acceptability in a care setting. No effects were predicted of age. These predictions were expressed as the following hypotheses:

H 2.4.1 The static and dynamic systems will be ranked most appropriate for a care scenario.

H 2.4.2 There will be no effect of age on the appropriateness for care rankings.

6.3.3 Research Question 3 – Subjective Workload

Research question 3 asked "*does the Dyna-Cue prototype create a higher workload compared to alternative reminder delivery methods?*" This was measured with the NASA-TLX assessment as defined in Section 6.2.5.

It was predicted that the dynamic condition will result in a smaller workload compared to the other conditions due to its use of low-salience interaction modalities and attempts to avoid overloading sensory channels. It is also predicted age will have an effect, with older participants reporting a higher subjective workload. These predictions are encapsulated in the following hypotheses:

H 3.1 The dynamic condition will result in a lower subjective workload than the other conditions for younger users.

H 3.2 The dynamic condition will result in a lower subjective workload than the other conditions for older users.

H 3.3 The older participants will report a higher subjective workload than younger participants.

6.4 Participants

Similarly to the study carried out in Chapter 4, two user groups were desired for the experiment: one younger, one older. The aim was to recruit 20 participants in each group with an even gender division. The younger user group was again limited to

18-30 years of age, and was recruited from students at the University. A total of 20 participants made up the younger group, of whom 10 were male and 10 were female. The mean age of the younger participant group was 20.5 ($\sigma = 1.6$).

The work carried out in Chapter 4 shows a significant effect of age in the older user group, which was defined as being 50 years old or more. To ensure that the 'older' age group accurately represented the older population, the age requirement for this study was raised so that all participants were over 60 years of age. Initially, the older user group was recruited from the same mailing list as the one used in Chapters 3 and 4. Unfortunately, the return rate for this experiment was extremely low, with only 6 participants recruited in this way; many of the replies from the mailing list were in their late 50s, and could not participate due to the new age requirement.

This issue was resolved by liaising with Prof. Vicki Hanson at the University of Dundee, who had access to a larger group of older participants. The remaining 14 participants were scheduled and the remainder of the study was carried out under laboratory conditions at the University of Dundee. Unfortunately, 4 participants did not show up as scheduled. However, it was possible to recruit additional participants at short notice from a technology education group run for older users at the University. As a result of the limited availability of participants, it was not possible to create a perfect gender balance. A total of 19 participants made up the older group, of whom 8 were male and 11 were female. The mean age of the older participant group was 69.3 ($\sigma = 7$).

Cognitive ability was not screened, in part because the experiment itself *is* a prospective memory test. However, one participant in the older group had significant issues with the experiment. While there were no outward signs of any prospective memory impairment, the participant became visibly distressed during the experiment and as such their participation was halted. The participant was paid for their time and their data destroyed. No other participants demonstrated these symptoms, but the event raised the question of whether participants' cognitive abilities should have been screened. This is discussed further in Section 6.7.

6.5 Procedure

All participants were given an initial demographic survey which included a sensory impairment self-assessment. No significant impairments were reported, although many older participants reported impairments that were corrected by hearing aids or spectacles. Participants were talked through the modified Hotel Test's primary activities, and the clock and secondary task buttons were demonstrated. Participants were given the opportunity to ask any questions about the tasks at this point.

Participants were then introduced to the notifications that they might receive during the experiment. To ensure understanding of the visual and auditory icons, participants were asked to try and interpret them without assistance. This was followed by a discussion and a 'retest' if the participant had difficulty interpreting them. While there were no issues with the visual icons, some participants had difficulty with the auditory icons. In particular, the chopping-board auditory icon for the 'cook' task was frequently misidentified as a door being knocked. None of the participants had any further recognition problems once the sound had been labelled and the retest administered. The tactile vibration was demonstrated and participants were informed that if they detected a vibration, they were to consider their schedule and current activity to interpret the meaning of the vibration. This is the same way participants were asked to respond to the beeps in the work of Manly *et al.* [133].

There were four experimental conditions in the experiment, each of which involved completing the modified Hotel Test once. The conditions were counter-balanced by gender and age to prevent ordering effects. At the start of each condition the type of assistance provided was explained so that participants fully understood the type of notifications to expect. Participants were also informed that the reminders would be delivered after a 3 second 'grace period'; therefore reminders would *only* be delivered if the participant was *already* late. This meant that participants could theoretically avoid reminders entirely by staying on schedule.[6]

At the start of each condition participants were given an instruction card with the times they were to press the 3 buttons, which they could keep at hand for reference. These instruction cards are provided in Appendix C.

During the experiment the participants were observed by the experimenter, who sat out of view behind the participant as shown in Figure 6.7b. When the participant started or stopped an activity they had to pull the relevant materials towards the center of the table, as shown in Figure 6.7a. The experimenter would then log the participant's activities using the control panel shown in Figure 6.7c. This proccess both logged

[6]In practice, only two participants were able to produce 'perfect scores' of 0, and both participants were only able to do so in one of their four conditions.

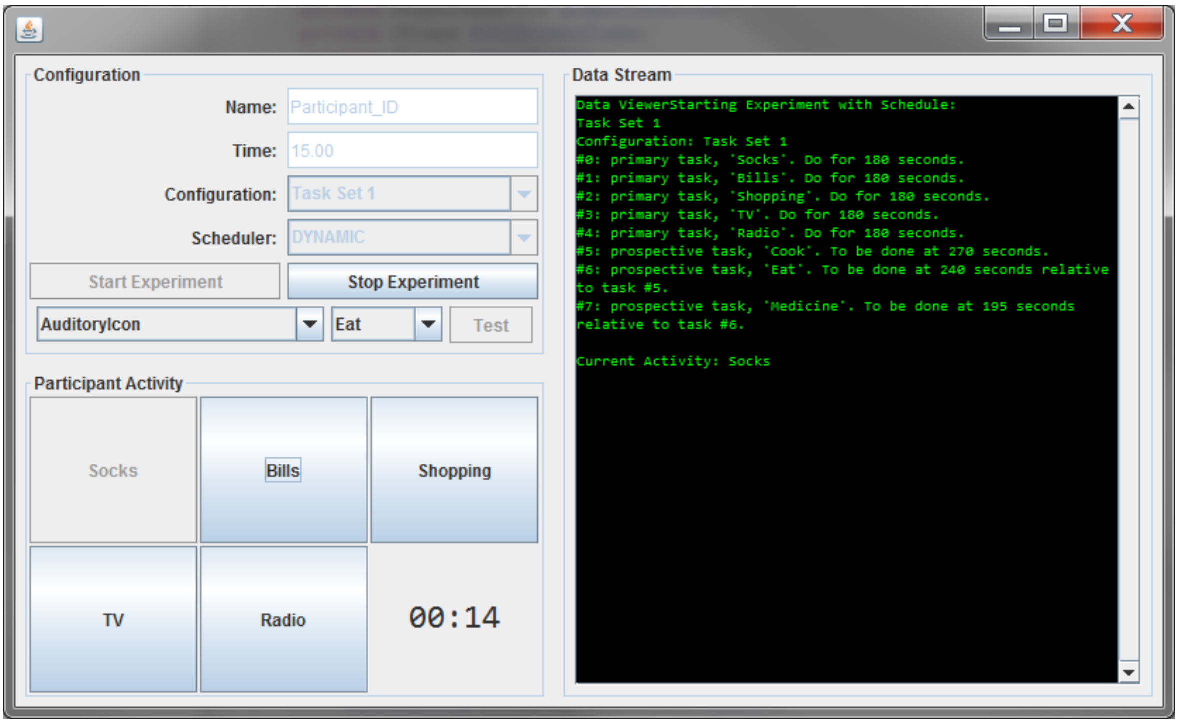
the participant's behavioural data and provided input to the Dyna-Cue scheduler as described in Section 5.4. Button data were recorded automatically and listed in the 'Data Stream' window shown to the right of Figure 6.7c. This allowed the experimenter to check in real-time that button presses had been acknowledged and to observe the delivery of tactile reminders, which allowed important observations to be made that are discussed in Section 6.7.

Between conditions the primary activities were reset using a new TV show, shopping list and collection of receipts. The paired socks were counted in full view of the participant to help prevent participant apathy. This seemed to be effective with many participants wanting to match more socks than their previous attempt. Participants were asked to complete the NASA-TLX assessment while the experiment was being reset.

When all conditions were completed participants were given the subjective questionnaire (see Section 6.2.5 and Appendix C) followed by a brief informal interview. The experiment lasted 80-90 minutes and participants were paid £10.

(a) Experimenter Position.

(b) Task signalling.



(c) Experimental Software Control Panel.

Figure 6.7: Photographs showing how the experimenter would observe the participant. Figure a shows how the experimenter will sit to the right, allowing them to watch the participant and record their actions. Figure b shows that the participant has pulled the remote control towards them. The experimenter would then press the 'TV' button (see Figure c) which would activate the TV.

6.6 Results

This section presents the results of the study, organised by the hypotheses given in Section 6.3.

6.6.1 Hypothesis 1.1 (Task Organisation)

Hypothesis 1.1 considered the effects of age and the type of assistance given on the Task Organisation Score (TOS). It is split into three sub-hypotheses examining age, assistance offered and any interactions between those two factors. A mixed-models general linear model (GLM) was constructed to test the hypotheses with age as a between-groups variable and the type of assistance offered as a repeated measures variable. An overview of the procedure used when carrying out GLM tests is provided in Appendix D.1. Mauchly's test was significant ($\chi^2(5) = 150.5,\ p < .05$), which signalled that the assumption of sphericity had been violated; therefore the model was corrected using the Greenhouse-Geisser method ($\epsilon = .37$). This data are shown in Figure 6.8.

Hypothesis 1.1.1 was "the dynamic and static conditions will both produce a lower TOS than the 'no assistance' condition, with the dynamic condition performing *at least as well as* the static system". The model showed that the type of assistance offered had a significant main effect on the TOS ($F(1.11,\ 41.21) = 13.32,\ p < .001,\ \eta^2 = .27$). *Post hoc* pairwise comparisons, shown in Table 6.2, showed that the only significant differences existed between the no reminders condition and the 3 conditions with reminders. The evidence supports the hypothesis, showing that both the static and dynamic reminders resulted in a considerable improvement to task organisation.

Hypothesis 1.1.2 was "the age of the participant will affect the TOS". The model found a significant main effect of age on TOS ($F(1,\ 37) = 50.65,\ p < .001,\ \eta^2 = .58$). This effect is very strong, as shown in Figure 6.8. The evidence supports the hypothesis that age is an important factor in determining organisation.

Hypothesis 1.1.3 was "there will be no interaction between modality and age on the TOS". However, the model found a significant interaction effect between age and assistance offered ($F(1.11,\ 41.21) = 5.83,\ p < .05,\ \eta^2 = .14$). Figure 6.8 suggests that the older participants benefited much more from the presence of reminders, with their performance almost matching the younger participants in the static condition. Table 6.3 shows the 95% confidence intervals for the model, which confirms that the older participants had lower performance in the 'no reminders' and dynamic conditions. The evidence in this case does not support the hypothesis, instead suggesting that the opposite is true.

For both the younger and older participants, all three forms of assistance improved their task organisation performance. The *post hoc* tests and the interaction effect suggest that the reminders conferred a greater performance benefit to the older users than it did for the younger users. The evidence shows that the Dyna-Cue system was effective at helping both younger and older users manage their time. However, both the static and random reminders were also shown to be effective. This finding will be discussed further in Section 6.7.

6.6.2 Hypothesis 2.1 (Difficulty)

Hypothesis 2.1 considered the effects of age and the type of assistance given on how difficult participants would consider the test. The data are shown in Figure 6.9, and the full analysis of the data can be found in Appendix C in Table C.6.

Hypothesis 2.1.1 was "the dynamic and static conditions will both be consistently ranked easier than the 'no assistance' condition, but will not be significantly different from each other". A non-parametric Friedman's ANOVA found a main effect of the condition on subjective difficulty for both younger ($\chi^2(3) = 21.3,\ p < .001$) and older ($\chi^2(3) = 14.5,\ p < .01$) participants. For the younger participants *post hoc* tests with Bonferroni corrections revealed a significant difference between the condition without reminders and both the dynamic ($p < .01$) and the static ($p < .001$) conditions. For older participants *post hoc* tests with Bonferroni corrections revealed the condition with static reminders was significantly different from the other conditions ($p < .05$), but no other differences were found. The evidence in this case shows that the dynamic and static conditions were considered easier than the condition without notifications as predicted, but only for younger participants. Older participants found the static condition easiest, with the dynamic condition showing no significant difference from the condition without reminders. The findings partly confirm the hypothesis.

Hypothesis 2.1.2 was 'the age of the participant will not affect the subjective difficulty rankings". This was tested by running a Mann-Whitney U test for each of the 4 conditions, then manually applying the Bonferroni correction. The results of this process are provided in Appendix C in Table C.6c. The analysis found no evidence that participant age affected subjective difficulty.

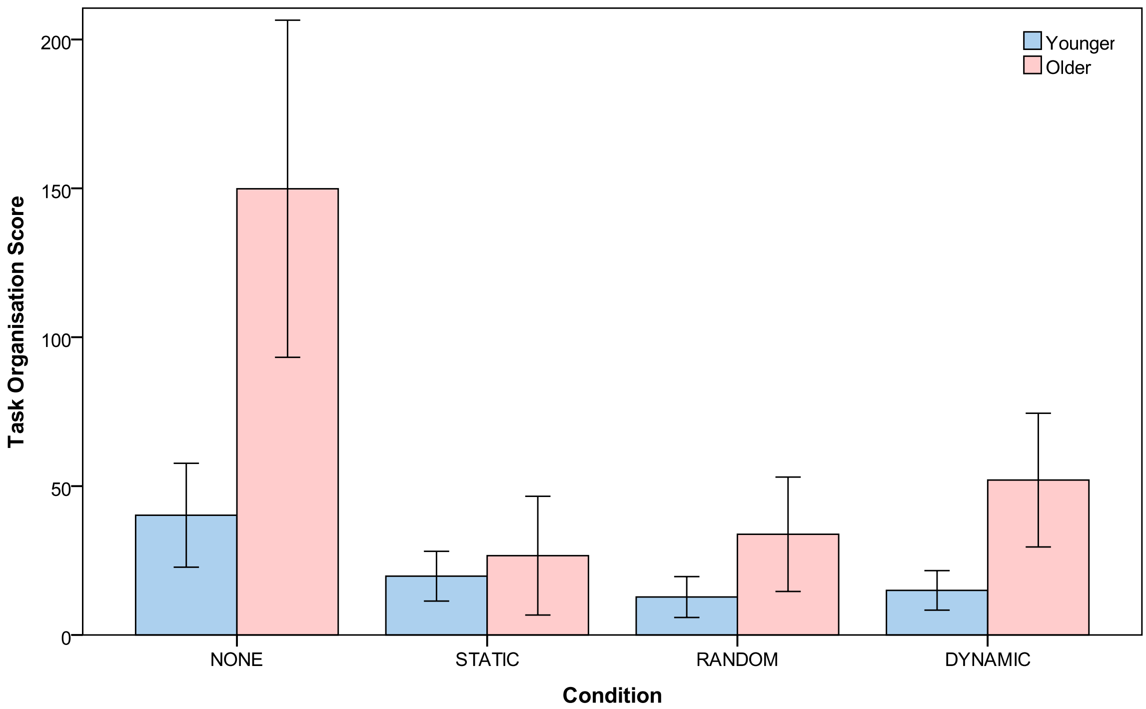
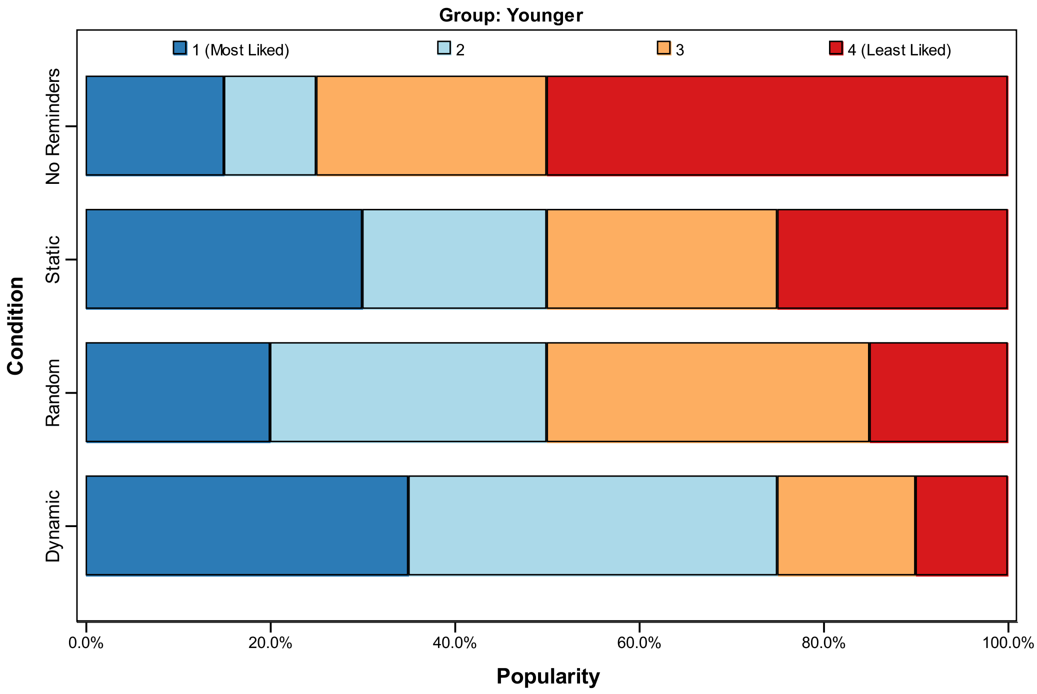


Figure 6.8: Graph showing task organisation scores by condition and age group. Error bars show 95% confidence intervals calculated using the Cousineau method as described in Appendix D.2.

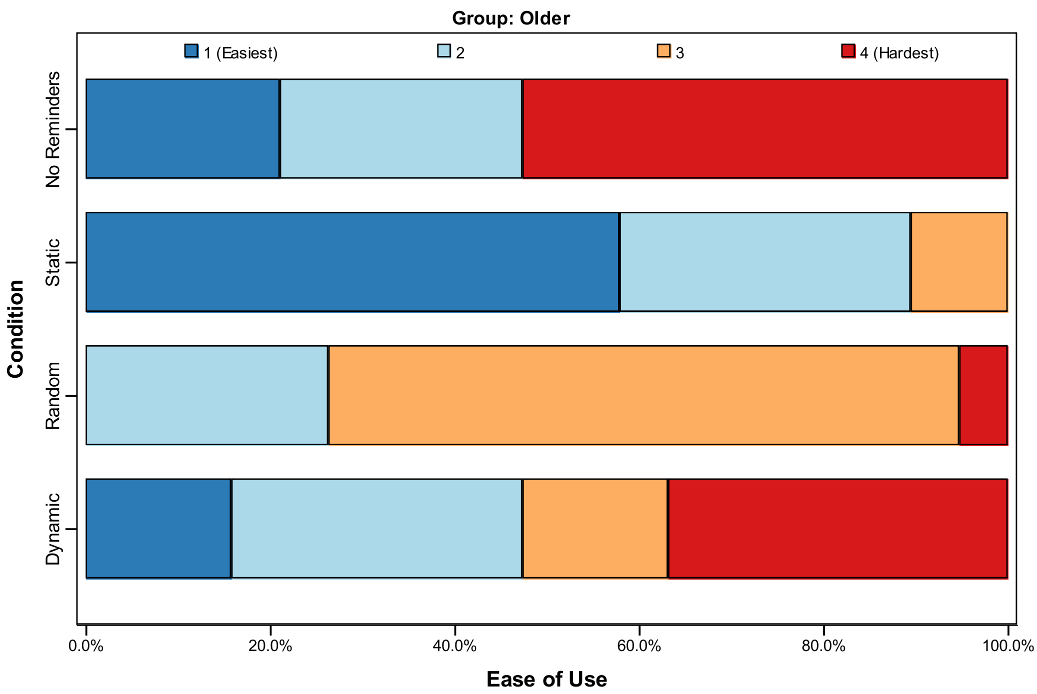
Table 6.2: Pairwise comparison of type of assistance provided and task organisation score (TOS).

	Mean	No Reminders	Static	Random	Dynamic
No Reminders	95.04		71.85^{**}	71.76^{**}	61.55^{*}
Static	23.19	-71.85^{**}		-0.09	-10.31
Random	23.28	-71.76^{**}	0.09		-10.21
Dynamic	33.49	-61.55^{*}	10.31	10.21	

Note: Table shows means and difference between means with significant p values noted. Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.



(a) Younger Participant Group ($N = 20$).



(b) Older Participant Group ($N = 19$).

Figure 6.9: Graphs showing the subjective difficulty of the conditions for the younger and older groups. The segments of each bar show the distribution of rankings for the conditions.

Table 6.3: 95% Confidence intervals for the interaction between age and type of assistance. Calculated by SPSS. $\overline{TOS}$ is the mean TOS, SE the standard error and CI the upper/lower 95% confidence intervals.

Condition	Younger				Older			
	$\overline{TOS}$	SE	CI_L	CI_U	$\overline{TOS}$	SE	CI_L	CI_U
No Reminders	40.20	26.93	−14.36	94.77	149.87	27.63	93.89	205.85
Static	19.74	5.38	8.85	30.63	26.63	5.52	15.46	37.81
Random	12.74	5.70	1.20	24.29	33.82	5.85	21.97	45.66
Dynamic	14.96	5.61	3.61	26.32	52.01	5.75	40.36	63.66

The results of the analysis demonstrate that the static and dynamic conditions were considered the easiest by younger participants, but the older participants preferred the static condition in general. This finding is discussed in Section 6.7.

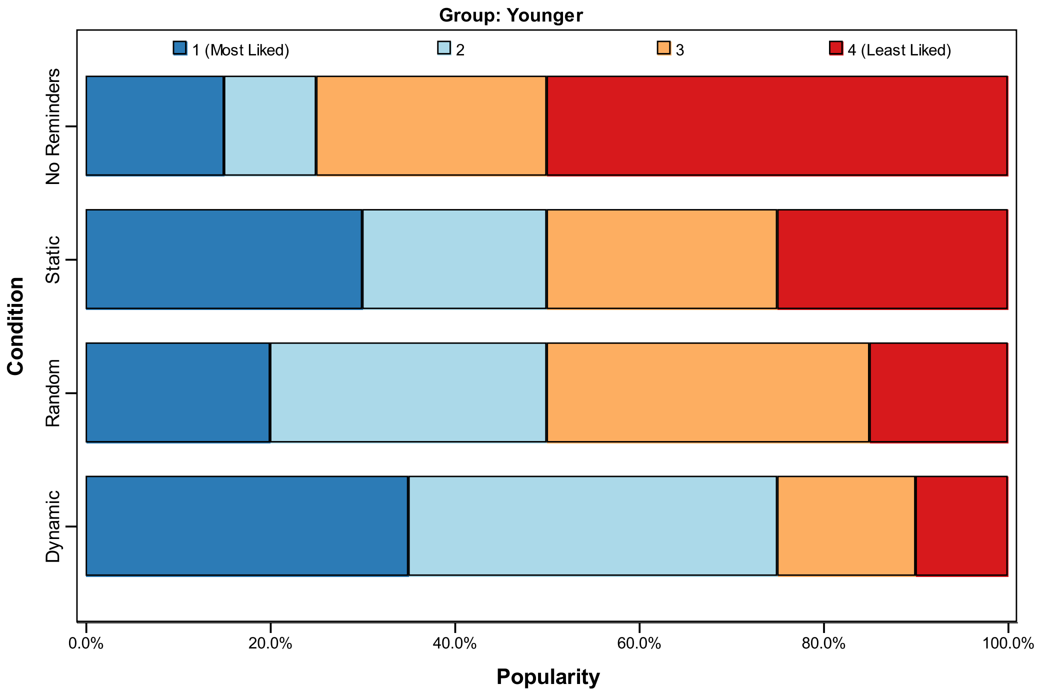
6.6.3 Hypothesis 2.2 (Preference)

Hypothesis 2.2 considered the effects of age and the type of assistance given on participant preferences. The data are shown in Figure 6.10, and the full analysis of the data can be found in Appendix C in Table C.7.

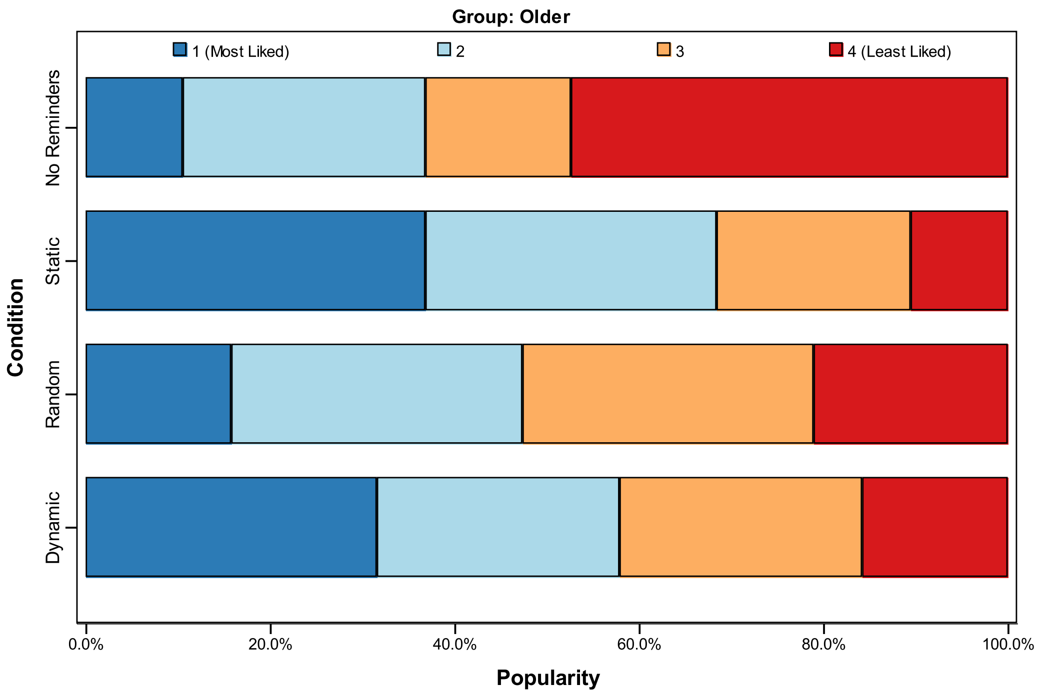
Hypothesis 2.2.1 was “the dynamic condition will be preferred by participants”. A non-parametric Friedman’s ANOVA found no effect of condition on participant preferences for either the younger group ($\chi^2(3) = 7.38,\ p = .06$) or the older group ($\chi^2(3) = 6.13,\ p = .11$). The results show a low level concordance: Kendall’s W was found to be .12 for younger participants and .11 for older participants, suggesting varying preferences. The evidence does not support the hypothesis, instead suggesting that there was no favourite form of assistance.

Hypothesis 2.2.2 was ‘age will not have an impact on preference”. This was tested by running a Mann-Whitney U test for each of the 4 conditions, then manually applying the Bonferroni correction. The results of this process are provided in Appendix C in Table C.7b. The analysis found no evidence that participant age affected subjective preference.

The results demonstrated that the experimental participants did not prefer the Dyna-Cue system as expected, instead showing a range of opinions. Post-experiment interviews provided an important insight into the diversity of preference scores, which is discussed further in Section 6.7.



(a) Younger Participant Group ($N = 20$).



(b) Older Participant Group ($N = 19$).

Figure 6.10: Graphs showing the subjective popularity of the conditions for the younger and older groups. The segments of each bar show the distribution of rankings for the conditions.

6.6.4 Hypothesis 2.3 (Appropriateness for Home)

Hypothesis 2.3 considered the effects of age and the type of assistance given on participant preferences. The data are shown in Figure 6.11, and the full analysis of the data can be found in Appendix C in Table C.8.

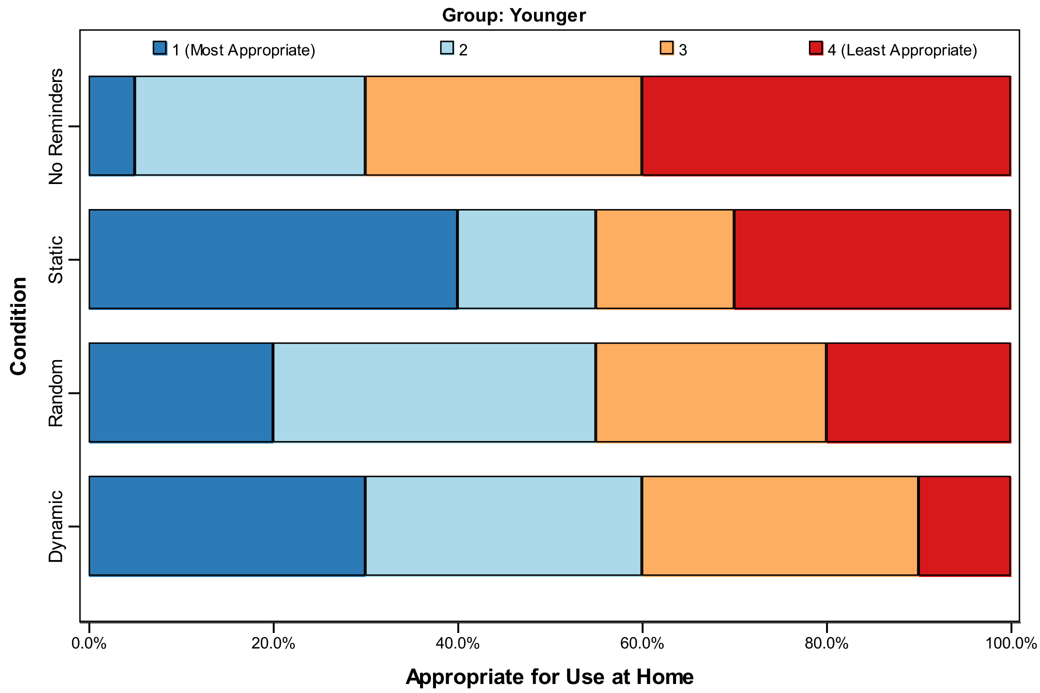
Hypothesis 2.3.1 was "the dynamic system would be considered most appropriate for the home by older participants". A non-parametric Friedman's ANOVA found a main effect of the condition on appropriateness for use at home for the older user group ($\chi^2(3) = 10.4,\ p < .05$), but not for the younger user group ($\chi^2(3) = 5.26,\ p = .15$). *Post hoc* pairwise comparisons with Bonferroni corrections were carried out on the older user data, which revealed that the only significant differences were between the condition without reminders and the static reminder condition ($p < .05$). These findings do not support the hypothesis.

Hypothesis 2.3.2 was 'there will be a difference in the subjective assessment of appropriateness for the home between the two age groups". This was tested by running a Mann-Whitney U test for each of the 4 conditions, then manually applying the Bonferroni correction. The results of this process are provided in Appendix C in Table C.8c. The analysis found no evidence that participant age affected perceived suitability for the home.

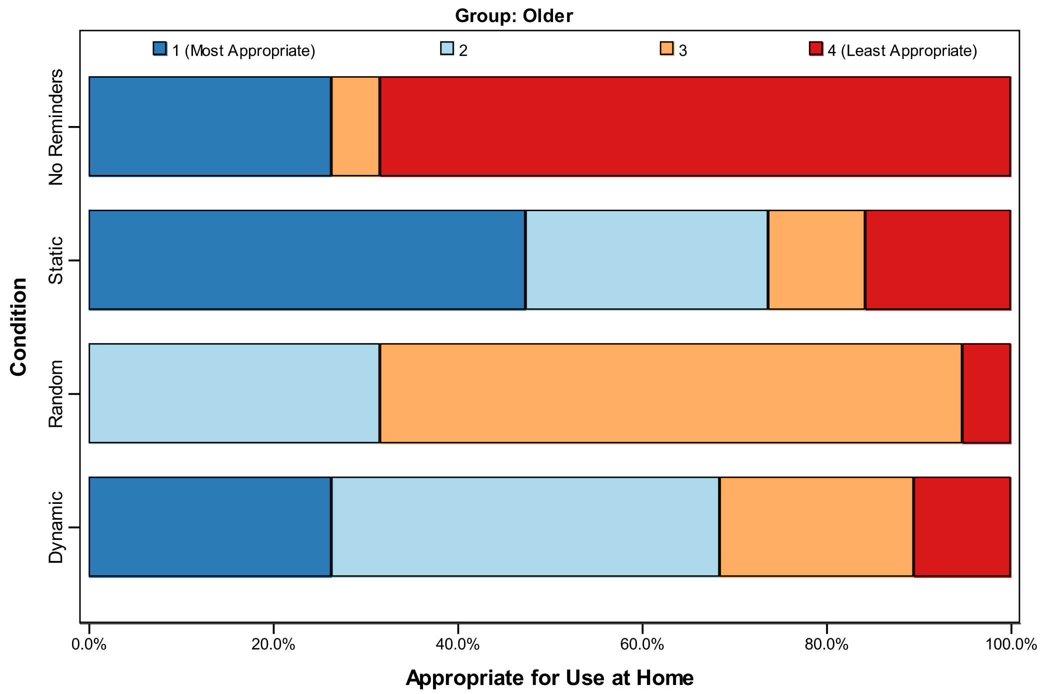
It was predicted in Section 6.3.2 that younger participants would not prefer a reminder system at all, as they are unlikely to have experienced any age-related prospective memory decline. A low degree of concordance was found for the younger participants ($W = .09$), suggesting considerable disagreement about which technology would be most appropriate for the home. The older participants were expected to prefer the subtlety of the Dyna-Cue system, but the evidence instead suggests that participants preferred the static system. Despite reaching significance, concordance for the older participants was low ($W = .18$), perhaps best highlighted by the fact that 5 participants thought no reminders would be *most appropriate*, while 13 considered that to be *least appropriate* (see Table C.4). This finding is discussed further in Section 6.7.

6.6.5 Hypothesis 2.4 (Appropriateness for Care)

Hypothesis 2.4 considered the effects of age and the type of assistance given on participant preferences. The data are shown in Figure 6.12, and the full analysis of the data can be found in Appendix C in Table C.9.

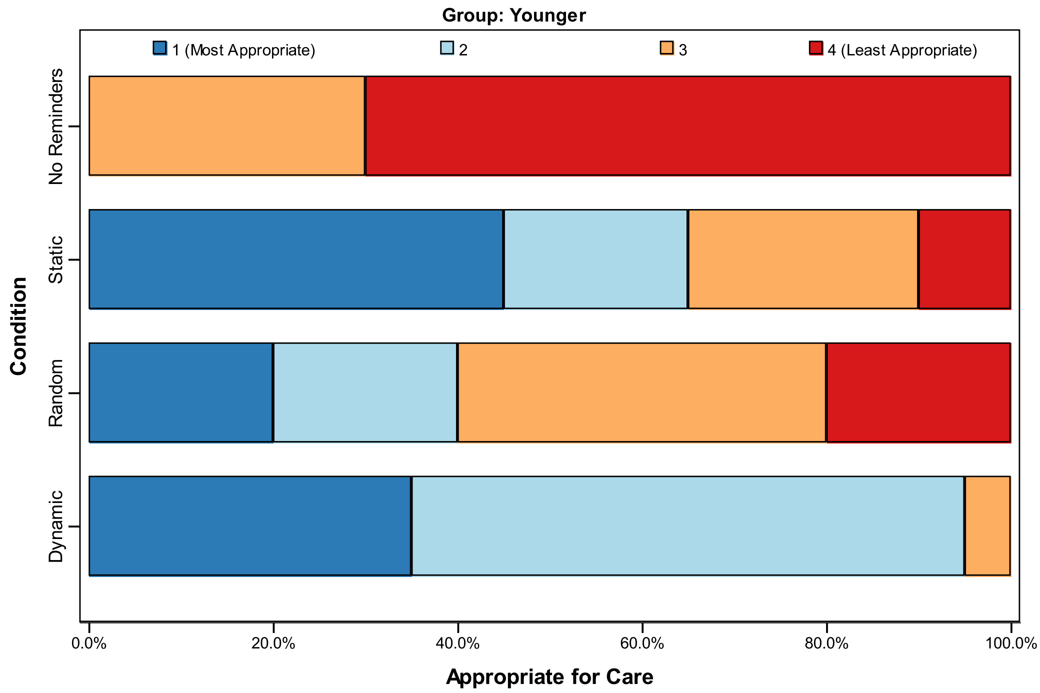


(a) Younger Participant Group ($N = 20$).

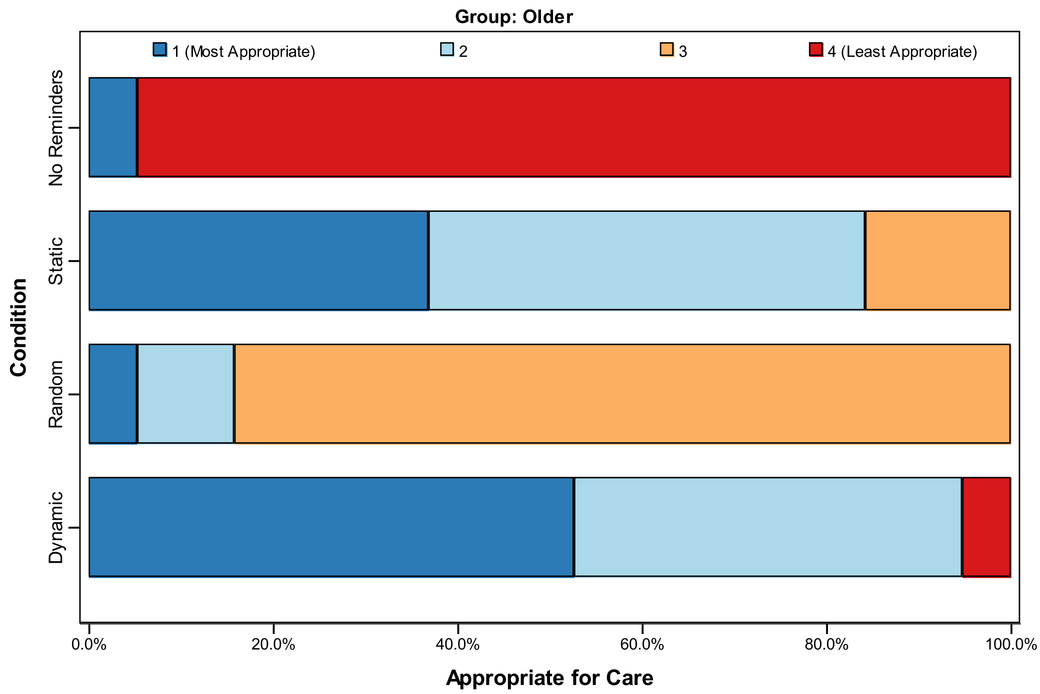


(b) Older Participant Group ($N = 19$).

Figure 6.11: Graphs showing the subjective 'appropriateness for the home' of the conditions for the younger and older groups. The segments of each bar show the distribution of rankings for the conditions.



(a) Younger Participant Group ($N = 20$).



(b) Older Participant Group ($N = 19$).

Figure 6.12: Graphs showing the subjective 'appropriateness for a care setting' of the conditions for the younger and older groups. The segments of each bar show the distribution of rankings for the conditions.

Hypothesis 2.4.1 was "the static and dynamic systems will be ranked most appropriate for a care scenario". A non-parametric Friedman's ANOVA found a main effect of the condition on appropriateness for use in a care setting for both younger ($\chi^2(3) = 28.1,\ p < .001$) and older ($\chi^2(3) = 36.9,\ p < .001$) participants. There was also a high level of agreement between participants as shown by Kendall's coefficient of concordance $w = .55$. *Post hoc* analysis of the younger participants revealed that the no-reminder condition was significantly different from all others, but the conditions with reminders were not significantly different. The older participants identified 3 groups; a dynamic-static group, a random group and a no-reminders group. This is shown in Table C.9b. The evidence in this case supports the hypothesis that the dynamic and static conditions would be considered most appropriate for care.

Hypothesis 2.4.2 was 'there will be no effect of age on the appropriateness for care rankings". This was tested by running a Mann-Whitney U test for each of the 4 conditions, then manually applying the Bonferroni correction. The results of this process are provided in Appendix C in Table C.9c. The analysis found no evidence that participant age affected perceived suitability for use in a care setting.

This result shows that the dynamic condition and the static condition were considered by participants to be highly appropriate for use in a care situation, while the other methods were considered less appropriate. This result is discussed further in Section 6.7.

6.6.6 Hypothesis 3.1 (Workload – Younger)

Hypothesis 3.1 was defined in Section 6.3.3 as "the dynamic condition will result in a lower subjective workload than the other conditions for younger users". Workload (WL) was measured using the NASA-TLX assessment, which uses 6 independent components to calculate an overall workload. As in Chapters 3 and 4, both the overall workload and individual elements were examined to provide an insight into the workload. The 6 workload component measures are Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Overall Performance (OP), Effort (EF) and Frustration (FR). The overall workload and its composition is shown in Figure 6.13.

Table 6.4 shows the result of a set of Friedman's ANOVAs testing for an effect of modality on NASA-TLX scores. The table shows that for younger participants WL was significantly affected by the type of assistance provided, as was MD and FR. *Post hoc* tests were carried out on those factors which are included in Appendix B. The *post hoc* analysis can be summarised as follows:

WL Significant differences existed only between the condition with no assistance and the conditions that provided assistance (Table C.1b).

MD Significant differences existed only between the condition with no assistance and the conditions that provided assistance (Table C.2b).

FR Significant differences only existed between the condition without assistance and the random condition. (Table C.7b).

The evidence in this case does not support the hypothesis. Table 6.4 shows that Kendall's W was low for the stats, suggesting low agreement between participants. This is discussed in Section 6.7.

6.6.7 Hypothesis 3.2 (Workload – Older)

Hypothesis 3.2 was defined in Section 6.3.3 as "the dynamic condition will result in a lower subjective workload than the other conditions for older users". The overall workload and its composition is shown in Figure 6.13. Table 6.5 shows the result of a set of Friedman's ANOVAs testing for an effect of modality on the NASA-TLX scores. The table shows that the only factor affected by the condition was PD. *Post hoc* tests were carried out on PD and are provided in Table C.3b. The *post hoc* tests revealed that the only significant differences lay between the dynamic and random conditions. As with the younger participants, the evidence in this case does not support the hypothesis. The implications of this finding are discussed in Section 6.7.

6.6.8 Hypothesis 3.3 (Workload – Between Groups)

Hypothesis 3.3 was defined in Section 6.3.3 as "older participants will report a higher subjective workload than younger participants". This was testing by taking the mean NASA-TLX scores for each participant and comparing the means using the independent-samples Mann-Whitney U test, the results of which are shown in Table 6.6. Graphs that directly compare workload components are provided in Appendix C.

Table 6.6 shows that the only significant differences between the groups were in the OP and FR components of the workload. The older participants appeared to rate their performance worse when compared to younger participants, which is highlighted by

Table 6.4: Analysis of repeated-measures TLX factors for the younger participant group.

TLX	χ^2	Kendall's W	N	df	p
Overall Workload	14.39	0.24	20	3	0.002**
Mental Demand	22.70	0.38	20	3	<.001***
Physical Demand	7.82	0.13	20	3	0.050
Temporal Demand	3.63	0.06	20	3	0.304
Overall Performance	5.16	0.09	20	3	0.160
Effort	5.92	0.10	20	3	0.116
Frustration	9.28	0.16	20	3	0.026*

Note: Table shows Freidman's two-way ANOVA calculated using SPSS. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$. *Post hoc* tests are provided in Appendix B.

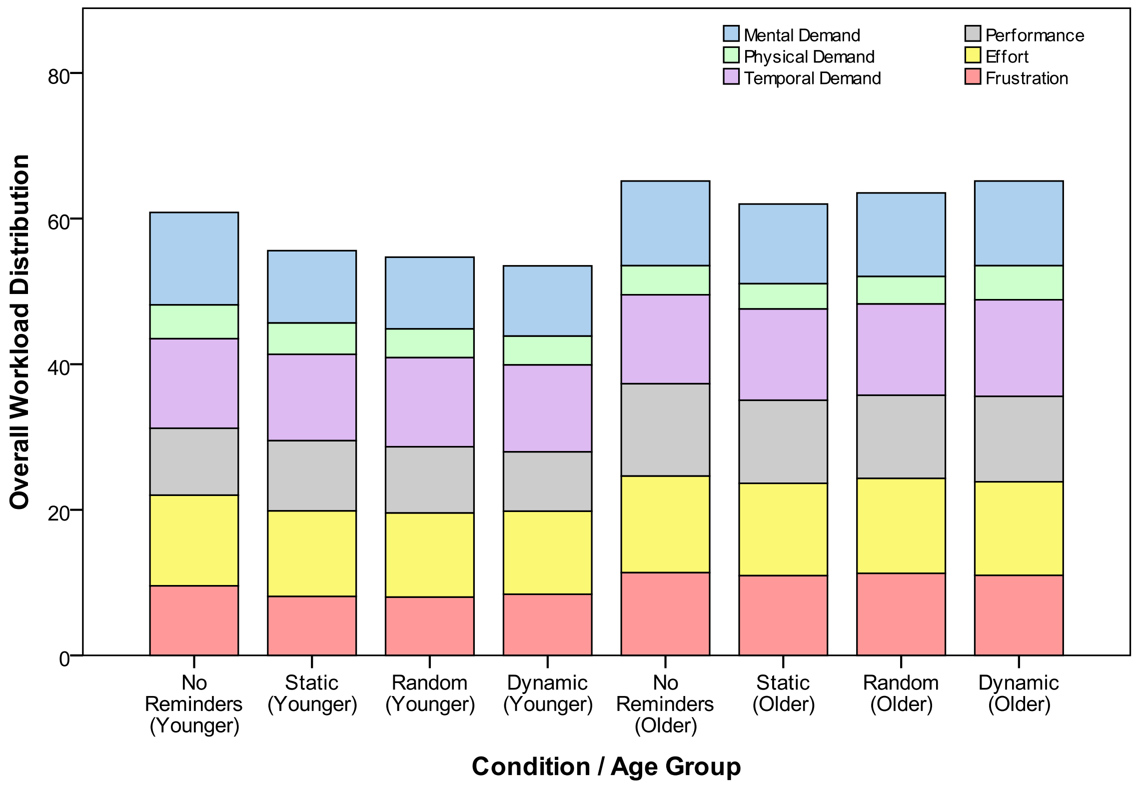


Figure 6.13: The relationship between modality and workload. The top of each bar shows overall workload.

Table 6.5: Analysis of repeated-measures TLX factors for the older participant group.

TLX	χ^2	Kendall's W	N	df	p
Overall Workload	0.62	0.01	19	3	0.891
Mental Demand	2.86	0.05	19	3	0.414
Physical Demand	7.86	0.14	19	3	0.049*
Temporal Demand	2.55	0.05	19	3	0.466
Overall Performance	1.97	0.04	19	3	0.578
Effort	0.79	0.01	19	3	0.851
Frustration	0.29	0.01	19	3	0.962

Note: Table shows Freidman's two-way ANOVA calculated using SPSS. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$. *Post hoc* tests are provided in Appendix B.

Table 6.6: Between-groups tests for the workload components.

TLX	U	Wilcoxon W	N	Z	p	p_{adj}	r
OW	248.00	438.00	36	1.63	0.103	0.052	0.27
MD	211.00	401.00	36	0.59	0.555	0.278	0.10
PD	191.00	381.00	36	0.028	0.977	0.489	0.00
TD	196.00	386.00	36	0.169	0.866	0.433	0.03
OP	267.00	457.00	36	2.166	0.030	0.015*	0.36
EF	226.50	416.50	36	1.027	0.304	0.152	0.17
FR	252.50	442.50	36	1.757	0.079	0.040*	0.29

Note: Table shows Mann-Whitney's U test and Wilcoxon's W. The tests were carried out using SPSS which produces two-tailed significance values (listed as p), but as the hypotheses were one-tailed the p values were adjusted to suit (listed as p_{adj}). Effect sizes are shown under r, calculated using the method suggested by Rosenthal [177] which is $r = \frac{Z}{\sqrt{N}}$. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.5. The same pattern is observed for the frustration measures, shown in Figure C.7.

Unlike the work presented in Chapters 3 and 4, the evidence in this case does not support the hypothesis and instead suggests that the overall workload is not affected by age. However, Table 6.6 shows that older participants perceived their performance to be lower and found the task more frustrating than younger participants. This will be discussed further in the following section.

6.7 Discussion

This section presents a discussion of the results of the experiment, starting with the task organisation data in Section 6.7.1. Three types of subjective feedback were gathered during the experiment: the subjective rankings, NASA-TLX subjective workload data and the post-experiment interviews. The rankings are discussed in Section 6.7.3 and workload in Section 6.7.2, with the interview feedback used to provide extra insights where appropriate. Finally, Section 6.7.4 provides a discussion on the validity of the experimental design, including subjective participant data that provided an insight into the ecological validity of the home-style tasks.

6.7.1 Task Organisation

Task organisation was assessed using the TOS measurement, calculated as described in Section 6.2.5. The results of the study showed that the Dyna-Cue system created a significant improvement over the control condition without reminders, as predicted. However, the results showed that the static and random conditions also resulted in a significant improvement, with none of the three systems being significantly better than the others for both younger and older users.

An interesting trend can be seen in Figure 6.8, which appears to show that the older users performed best in the static condition, getting worse in the random condition, and worst in dynamic condition. Table 6.3 confirms that performance in the dynamic condition was significantly lower than in the static condition, although it remained significantly better than the condition without reminders. This was surprising given that a similar trend was not observed for the younger participants.

As noted in Section 6.5, the experiment included a data viewer (see Figure 6.7) that allowed for real-time monitoring of the system. This allowed the experimenter to observe when notification delivery started and stopped, and allowed an important observation to be made: older participants frequently showed no sign that they received tactile notifications, which was not true for younger participants. The tactile device (a Nexus One android phone connected via Bluetooth) was tested several times as a result, but was found to be working as expected. Any issues communicating with the phone would have raised an error message visible to the experimenter. The 'tactile lag' reported in Chapters 3 and 4 was also unlikely to be responsible, as a similar effect was not observed in younger participants.

This suggests that the older participants frequently missed the tactile notifications. Using the data from Figure 5.4, the Dyna-Cue system would choose a solo tactile reminder 50% of the time for a low-importance reminder. Reminder re-delivery took 5 seconds, at which point the urgency score would be increased and the delivery method selection reconsidered (as discussed in Chapter 5). Hypothetically, if 4 'change task' reminders were delivered, with a 50% likelihood of being delivered solely via tactile, then it would suggest that the TOS for older participants in the dynamic condition would increase by around 10 seconds between the dynamic and static conditions. The actual difference is much larger than this at 25.38 seconds. However, the ineffectiveness of tactile reminders for older participants would also imply that the data used to define the behaviour of the Dyna-Cue system (Figure 5.2) was incorrect; the salience value of the tactile condition should have been lower for most of the older participants. This would have had a knock-on effect on the Dyna-Cue system's ability to make decisions. The implications of this are discussed further in Section 6.9.

Although the Dyna-Cue system failed to match the performance of the static system for older people, all three systems led to much better performance compared to the condition without reminders. The younger participants demonstrated the expected trend, with the Dyna-Cue system being as effective as more intrusive methods. However, the aim of the Dyna-Cue system was to be as effective as those technologies while being more appropriate.

The first research question asked *"how does the Dyna-Cue system compare to other forms of reminder technology with respect to helping people to manage time and organise activities?"*. The study has shown that Dyna-Cue is effective at helping people to manage their time and activities, but that other forms of technology were also effective. For older people, the configuration of the Dyna-Cue system used in the study was not as effective as consistent text and speech static notifications. However, it is likely that another configuration would allow the Dyna-Cue system to match the static system, as it did with younger participants. This could possibly be achieved by using different modalities or by reconfiguring existing modalities.

6.7.2 Subjective Rankings

The participants were asked to rank the conditions in terms of difficulty, preference, appropriateness for the home and appropriateness for care. The difficulty rankings showed that the 'no reminders' condition was the lowest, with the static condition generally voted easiest. For the younger participants the dynamic condition performed well compared to the static condition, however for older participants it was rated in a similar pattern compared to the no reminder condition (Figure 6.9). This reflects the performance measures, which were discussed in Section 6.7.1. One interesting insight provided by the difficulty rankings was that there was no correlation between found between the ordering of the conditions and difficulty rankings ($\rho = -0.14, p = .8$), which was interesting as the Hotel Test is known to have a learning effect [133]. The lack of correlation suggests that participants were able to answer the post-experiment survey without being unduly influenced by the order of the conditions.

Preference was one area where the Dyna-Cue prototype was expected to perform well, yet the results showed no preference for any of the conditions for either the younger or the older participants nor between the two groups. Figure 6.10 highlights this well, showing the rankings were comparable for both the younger and older participants. This distribution suggests a high level of variance between participants, which is highlighted by the low Kendall's concordance values of $W = .12$ for younger participants and $W = .11$ for older participants.

The post-experiment interviews provided an insight into the reasons behind the choices participants made. Some participants said they would not want or need reminder technology, and that it would annoy them at home. This was a common sentiment amongst the younger participants. The older participants were more open to the idea of reminders at home but favoured different methods. Some of the comments by participants included:

- "Static notifications would be same every time, so I would know to look out for them."

- "(static reminders) are really clear, so I know what they mean."
- "Random notifications would stop me from getting used to the notifications, so I wouldn't subconsciously ignore them."
- "I like the idea that the dynamic notifications would get more demanding (salient) if I didn't respond to it."
- (on the dynamic and random reminders) "I like the idea of having all the different ways to send messages, it means I'm more likely to notice them.
- "(static reminders) would annoy me at home, they are too obnoxious."
- (on random and dynamic reminders) "It is too hard to know what they mean, I think they would become confusing."

These comments highlight the importance of personal preference when designing technology for the home, which has been highlighted by other researchers [145, 207].

Appropriateness for the home and care were the final two questions on the ranking survey. In the appropriateness for the home rankings, the only significant difference found was between the 'no reminders' condition and the static condition for the older group, and no effect was found for age. It is likely that the reasoning behind the participant responses is linked to their preferences. The older participants in particular would often use their own home or daily activities as examples in the post-experiment interview.

Appropriateness for care was an important question for the evaluation of the Dyna-Cue prototype. The younger participants considered all the reminder systems to be more appropriate than no assistance at all, while the older participants considered the static and dynamic systems to be most appropriate.

The second research question was "*does the Dyna-Cue prototype produce more appropriate interactions compared to alternative reminder delivery methods?*". The results of the study show that this is not the case; there is in fact a large amount of disagreement between participants regarding which of the methods was the most appropriate.

An important point is that the Dyna-Cue technology presented here is intended to be customisable to allow for personal preferences (*e.g.* by adding and removing modalities or changing the base scores used to make selections). As noted in Section 6.2.3, the static and dynamic systems used in the study were both created by customising the Dyna-Cue prototype slightly. In practice this means that the participants who preferred the static and random systems could be satisfied using the a reconfigured Dyna-Cue prototype. However, a static piece of technology could not be reconfigured in the same way. This suggests that the next version of the Dyna-Cue prototype should make user preferences central to decision making, possibly providing several different modes of operation (*e.g.* static, random, dynamic) to suit the user. This is discussed further in Section 6.9.

6.7.3 Subjective Workload

Like the work carried out in Chapters 3 and 4, the NASA-TLX data demonstrated a difference between the control condition and the experimental conditions, but no consistent differences were found between the experimental conditions. Comparing the age groups however revealed a significant difference in perceived performance and frustration, with older participants rating their performance lower and their frustration higher. This is interesting when taking into consideration the fact that no effect of was found of assistance on workload for the older participants, *i.e.* the older participants rated their frustration and performance higher across all the conditions, as shown in Figures C.5 and C.7. This is not in alignment with their actual performance, which increased dramatically with the assistance of notifications as shown in Section 6.6.1. One possible reason for this is that the tasks themselves, which could not be completed in the allotted time, interfered with the NASA-TLX ratings.

The third research question was "*does the Dyna-Cue prototype create a higher workload compared to alternative reminder delivery methods?*". The evidence suggests that none of the systems produced a higher workload than the others, including the condition without reminders.

6.7.4 Reflection on Experimental Design

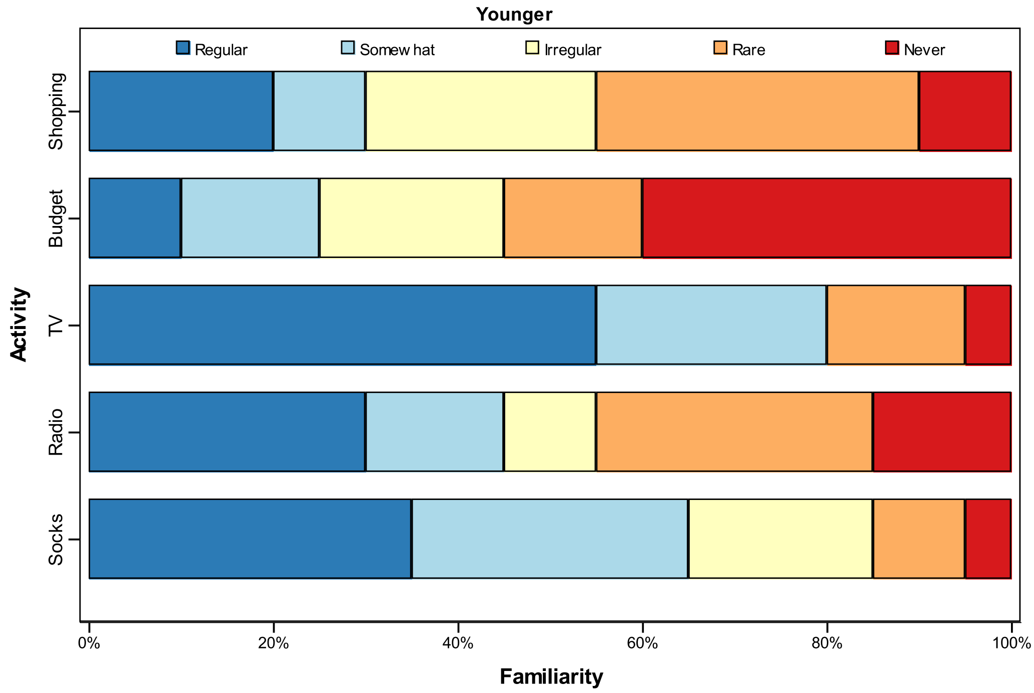
As part of the study participants were asked to consider whether each of the 5 primary activities represented a regular household activity. This was measured on a 5-point scale ranging from "regular household activity" to "never carried out at home", provided in Appendix C. The results of these questions are shown in Figure 6.14, which revealed that the Laundry and TV tasks represented a regular household activity for most of the participants. When asked to describe a similar task for the Shopping activity, most participants specified Internet shopping. Listening to MP3s or CDs was suggested as an alternative to the radio task. This shows that the experimental activities were not unlike activities that participants would regularly carry out in their own homes, as the ‘similar activities’ tended to involve the same modalities and mental processes.

The data also revealed the effect of age. The older group was much more likely to listen to the radio and create a household budget than the younger participants. For the younger participants, the Budget task was the only task for which a suitable analogue could not be found. Apart from the budgeting task for younger participants, the data suggested that all of the activities represented (or were analogous to) activities that the participants would often carry out at home, suggesting that the design of the study was ecologically valid.

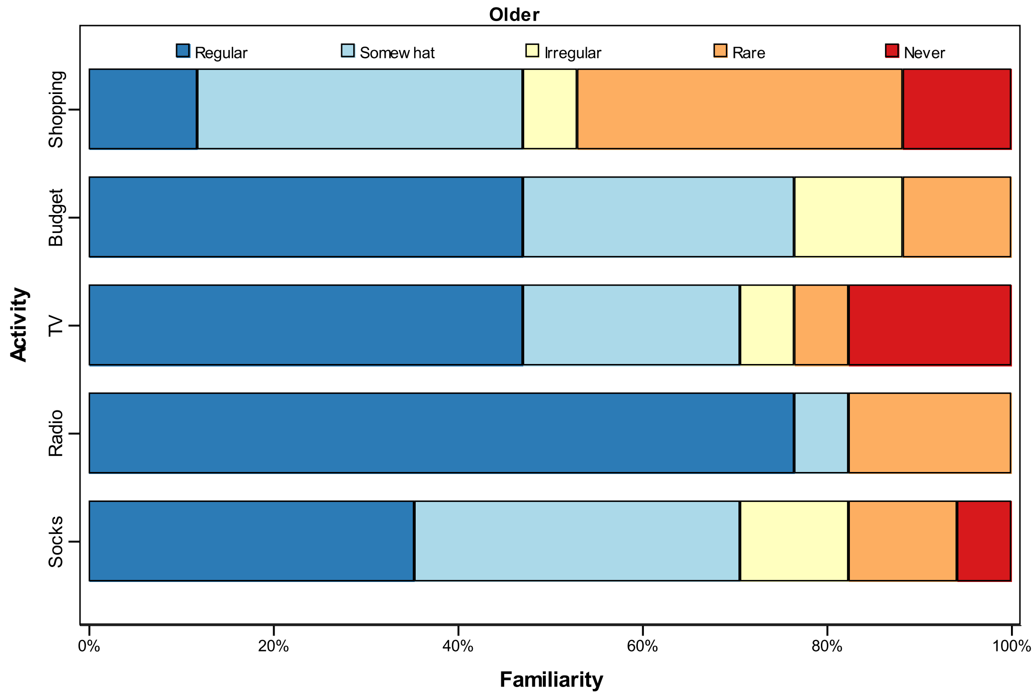
In Chapters 3 and 4, the experimental design kept in mind that older participants were less likely to be experienced at using computer technology. This ideal was also kept in mind for this experiment; as a result, the experiment was designed in such a way that the most complex technological interaction was pressing the large plastic buttons shown in Figure 6.3. As a result any effect from computer experience on the results would be minimal, suggesting that the results showing significant differences between age groups were influenced by other factors.

The visual notifications shared the TV space by ’shrinking’ the video feed, creating a black space at the bottom of the screen in which the text and pictographic messages were displayed. It might also be possible to overlay such images onto the screen without resizing, which might reduce the salience of those messages. The video feed in the experiment was created using a Java library called vlcj[7] which used an external vlc media player to render the video in a Java Swing component. Instead of attempting to draw additional elements on the video feed, a set area was defined at the bottom of the screen for visual notifications. The video ’shrinking’ would only effect the results during the TV task, with minimal impact on the outcome of the experiment. However, this may represent an interesting area for further study to ensure that the salience values for those modalities are correct.

[7] vlcj, https://code.google.com/p/vlcj/.



(a) Younger Participant Group ($N = 20$).



(b) Older Participant Group ($N = 19$).

Figure 6.14: Graphs showing the subjective familiarity with the home-style tasks for the younger and older groups.

6.8 Guidelines

The results of this study provided some important insights into the value of reminder technology, and in particular the properties of dynamic multimodal technology such as the Dyna-Cue prototype. Based on the results of the study presented here, the following guidelines were derived:

- Well-timed reminders can be highly effective at helping to manage time and task management regardless of the modality used. This is particularly true for older participants.
- Understanding the properties of the modalities is key to making a dynamic multimodal reminder system. In this case, the tactile messages were frequently missed by older participants, which was considered to be responsible for the performance difference between the static and dynamic systems for older participants.
- A dynamic reminder system might make more intelligent scheduling decisions, but personal preference plays the pivotal role in determining appropriateness for the home and care. A dynamic system would need to provide significant customisation options to ensure that it could satisfy the user's needs.
- The most appropriate methods were the static and dynamic methods. This suggests that many participants considered dynamic system to be highly appropriate. As the dynamic technology includes the static and random alternatives demonstrated here, more technology should be dynamic to provide more options to the end user. An important part of this is including multiple modalities.

6.9 Reflection on Dyna-Cue

Chapter 5 introduced a prototype multimodal home reminder system called Dyna-Cue, which was assessed in Chapter 6 against three alternatives: no assistance at all, a static system which always used text and speech, and a system which picked a modality at random. The subjective feedback provided in Section 6.6 showed that in both the younger and older groups there were several participants who preferred the random or static notifications over the Dyna-Cue system. The results suggested that preferences were independent of effectiveness and task ordering, similar to the findings of McGee-Lennon *et al.* [146]. The Dyna-Cue system was built on the assumption that more considerate technology would be preferred by participants [70, 203, 202]. However, the post-experiment interviews made it clear that some participants preferred *less* considerate technology in a home-care context. It is possible that the level of ecological validity reached by the experimental design was not sufficient to gather accurate subjective feedback on this matter, as the participants did not have to live with the technology in their own homes. More accurate data could only be gathered with longitudinal studies.

The Dyna-Cue prototype provided a new model for multimodal interaction which selected what it considered to be the most appropriate interaction modality based on several factors. Preference was not one of those factors, but in Chapter 5 it was suggested that preference could be used as a baseline score much in the same way as the salience and effectiveness scores. That is, the modality selected would also be influenced by preference scores.

The study suggested that this approach would not be sufficient to take preference into account. Instead, the logic of the system itself should be customisable, so that users can choose whether they will use Dyna-Cue-style notifications or static notifications. New work would be needed to find ways of allowing the end user and stakeholders to program such technology.

While the Dyna-Cue prototype was effective at aiding task organisation, it did not match the performance of the static reminders for older participants. In Section 6.7.3 it was suggested that the tactile modality could be the source of the performance drop as it did not seem to be salient enough for the older participant group. Figure 5.4 showed how the modalities would be scored based on the ratings of the delivery methods and the current environmental situation. If the salience value of tactile notifications was lowered from 2 (medium salience) to 1 (low salience), then the chance of receiving a tactile notification for a low-priority message would actually increase from 50% to 100%. The chance of getting a tactile or part-tactile notification for a medium-priority message[8] increased from 20% to 40%. This is the expected behaviour and was based on the work of Vastenburg *et al.* [202], who found that home-based reminders would be more appropriate if salience was reduced for low-priority reminders. By reducing the salience of the tactile modality, it would be used more often in for lower-value notifications. Vastenburg *et al.* noted that participants did not care if they did not

[8]Note that the non-cued prospective memory events (*i.e.* cook, eat and take medicine) would enter the system with an importance of 2.

receive low-priority notifications. In this context, participants were attempting to complete tasks against the clock and as such were likely to care more about receiving notifications. This provides further evidence that longitudinal studies would be needed in the home to fully understand how appropriate the Dyna-Cue system is.

An alternative perspective on the low salience of the tactile notifications would be to treat it as an example of how to deal with sensory impairment. If participants do not notice the tactile vibrations at all when it was working as intended, then communication in that channel is significantly impaired. This could be dealt with in several ways.

The modality which isn't working as expected could be removed from the Dyna-Cue system. The prototype required all modalities to be present when the system boots, but further work was carried out using Open Services Gateway Initiative (OSGI), which would allow modalities to be added and removed at runtime. Without a system reebot, the system would remain functional as hardware is added and removed. In addition, OSGI provides a mechanism through which new hardware could be developed to work with the existing system, allowing home care technology to 'evolve over time' in the manner suggested by Edwards & Grinter [57]. OSGI has already been used in this way by Maternaghan [135] in the HOMER system.

The system could track the effectiveness of sensory channels and adjust effectiveness or salience scores[9] in real time. This would also provide an interesting opportunity for monitoring the user's behaviour, *e.g.* a drop in the effectiveness of speech communication could be a sign of worsening aural impairment. Taken to extremes, a sudden drop in the effectiveness of a delivery method could be taken as a sign of stroke. In Section 2.2, existing home care technology was criticised for failing to interact with the user, instead focussing around monitoring. Strokes can be difficult to detect, and it is conceivable that technology which interacts with the user will be able to raise the alarm quicker than technology that waits for users to deviate from their normal routines.

Another observation, true for both younger and older participants, was that the visual notifications would change in salience depending on the user's activities. If the user was engaged in the Laundry task, they were likely to miss a notification on the TV as their attention was focussed elsewhere. Alternatively, if they were watching the TV then the sudden adjustment of the picture to make space for the message (see Figure 6.4) would be highly salient.

[9] This could be done across sensory channels or for each delivery method independently. Additional research would be needed to understand which of these models is the most appropriate.

If the user's activities are known it could be used to further improve the appropriateness of the Dyna-Cue prototypes scheduling decisions. The best example of this would be increasing the salience scores of visual methods when they have the user's attention, and decreasing their salience when they do not. Achieving this would require real-time knowledge of the user's activities.

The Dyna-Cue system only considered the status of the environment when making scheduling decisions, but it was aware of the user's activity due to the Wizard-of-Oz approach taken. In the real world, identifying the attentional focus of the user could be achieved in several ways, for example:

- Home automation systems such as X10 or Plugwise[10] could be used to identify the appliances which are currently turned on in the home. This would provide an output similar to the output from the Wizard-of-Oz control panel of the Dyna-Cue prototype, *i.e.* the current task. Interpreting that data would require the CI (see Section 5.4) has access to a database of task properties, as was done in the prototype.

- A device like the Microsoft Kinect could be used to track the user around the home. This would be more intrusive than monitoring individual appliances but would provide more fine-grained information about the user's activities. An interesting advantage of this approach would be the ability to deal with multi-user environments, as the Kinect would be able to identify different users based on their morphology [130]. This could also theoretically be used to detect social situations,
sensitive data.

Horvitz [94] introduced the idea of considering attention when making scheduling decisions, which was built upon by Iqbal & Bailey [103, 104] into the concept of breakpoints: moments between activities that are suitable points for interruption. The Dyna-Cue prototype did not take breakpoints into account, but during the pilot stage a 'grace period' was implemented as the system would often attempt to notify users as they were in the process of switching tasks.[11] Expanding the CI would be an important step in improving the Dyna-Cue system and would be required for longitudinal studies to take place.

[10]Plugwise, http://www.plugwise.com/idplugtype-g/.
[11]This would also happen for the static and random variants of the prototype.

Section 2.1.1 notes that with older users, cognitive decline is a significant issue. This issue arose during the experiment when one participant became distressed and the experiment was stopped (see Section 6.4). Although the Dyna-Cue system is capable of switching between modalities to account for sensory impairment simply by changing variables or adding/removing modalities, no clear provisions are made for cognitive decline. Cognitive decline takes many forms including various types of memory loss, and as such compensating for such changes is much more difficult than compensating for (relatively simple) sensory decline. Customising a system for cognitive decline would require a clearer focus on the individual's unique needs; from a technology perspective, the best that could be done is to make this process simple and easy. Although only a prototype, Dyna-Cue provides the ability to change how certain modalities are used (*e.g.* the message that is played by a speech reminder), and it is this type of technology that should be used when trying to tailor a system for a user with cognitive decline.

6.10 Conclusions

Thesis Question 4 considered if home care technology could be made more effective and appropriate through the ability to dynamically select from multiple forms of interaction. This chapter presented an ecologically valid method for assessing the Dyna-Cue prototype and a user study that demonstrated the benefits of reminder technology. While the Dyna-Cue prototype was shown to be effective at helping participants to manage their time and activities, the alternative methods tested also performed well, with the static system sometimes outperforming Dyna-Cue. Subjective feedback revealed a wide range of opinions regarding which method would be most appropriate in a home or care setting.

This result would appear to suggest that dynamic multimodal methods would not particularly improve the effectiveness and appropriateness of home care technology. However, one of the key aspects of the Dyna-Cue system is an ability to encapsulate a static system; *i.e.*, Dyna-Cue contains both a static and random reminder system. The results of the study suggest that participant preferences were more complex than expected, and that the level of dynamism provided by Dyna-Cue's base behaviour model is not sufficient to satisfy the expectations of the older users. To meet those needs, Dyna-Cue would need to be expanded on to allow more complex behavioural configurations, including the ability to customise the way modalities are selected.

More work would also be needed to fully understand if the lower performance of the Dyna-Cue system for older participants was due to the issue with tactile notifications. Longitudinal studies in the home of end users would be crucial in fully understanding the appropriateness and long-term benefits of using this type of technology. Future work should also consider ways to apply such dynamic technology to individuals with severe sensory impairments, and to find ways for stakeholder configuration of potentially complex dynamic systems such as the Dyna-Cue prototype.

In conclusion, the study has shown that dynamic multimodal technology can improve the effectiveness and appropriateness of home care technology. While the Dyna-Cue prototype's standard behaviour was not favoured by all participants, the alternative static and random systems were preferred by some. This suggests that a higher level of dynamism is needed to maximise effectiveness and appropriateness. That level of dynamism would not be attainable without multiple modalities and the technology within Dyna-Cue that allows it to fully utilise them. While further work needs to be carried out, the results of the study show that home care technology can be improved by the ability to dynamically select from multiple forms of interaction.

Chapter 7

Discussion & Conclusion

In Chapter 1, the aims of the thesis were set out as four thesis questions. Those questions were:

Thesis Question 1:
Which forms of interaction are appropriate for use in a home care system?

Thesis Question 2:
How do different forms of notification delivery affect users?

Thesis Question 3:
How can home reminder technology be designed to best utilise multiple types of interaction?

Thesis Question 4:
Can home reminder technology be made more effective and appropriate by providing it with the ability to dynamically select from multiple forms of interaction?

Thesis Question 1 was addressed in Chapter 3, which concluded that the range of interaction modalities suitable for use in the home was much wider than expected. Chapter 4 expanded on that work to address Thesis Question 2, gathering useful performance information about a range of modalities and identifying suitable application areas. Thesis Question 3 was addressed in Chapter 5, which combined the data gathered in Chapters 3 and 4 with existing work to create a dynamic multimodal reminder system called Dyna-Cue. Thesis Question 4 was addressed in 6, which used the Dyna-Cue system to explore the differences between different forms of reminder delivery mechanisms, and concluded that dynamic multimodal technology can improve the effectiveness and appropriateness of home care technology.

This section presents the conclusion of the thesis. A summary of the thesis is given in Section 7.1, followed by a discussion of the primary contributions of the work in Section 7.2. Section 7.3 discusses some of the limitations of the work. Section 7.4 discusses the future work that could be built on this thesis before Section 7.5 presents the final conclusion.

7.1 Thesis Summary & Main Contributions

Chapter 1 introduced home care technology and argued that improved home care technology is needed for ethical and economic reasons. One path to improve multimodal interaction would be the application of multiple modalities to make the technology more acceptable and versatile. Four high-level thesis questions were set out, asking *(1)* which modalities are suitable for use in a home care system, *(2)* what are the properties and application areas for those modalities, *(3)* how can technology fully utilise multiple modalities, and *(4)* would dynamic multimodal technology improve the effectiveness and appropriateness of home care technology.

Chapter 2 reviewed the relevant literature. This started by considering the requirements for home care technology, followed by an examination of home care projects and products in research and industry. This revealed that industry and research are generally working in different directions; industry projects were criticised for failing to interact with the user, while research projects tended to be impractical 'home of the future' proof-of-concept projects. The literature review proceeded to analyse reminder technology, including the psychology of interruptions and how different modalities have been used to deliver reminders. This revealed that while there is a lot of research on individual modalities, there is little existing work that provides a comprehensive comparison of the different modalities (or a taxonomy of their properties). It was also shown that while work has been carried out into dynamically managing *when* to deliver a message [94, 104], few people have considered managing *how* to interact [171, 202]. These represent important research gaps that need to be explored if multimodal interaction is to be applied to home care technology.

Chapter 3 presented an exploratory study that included eight modalities: text, pictograms, abstract-visual messages, speech, auditory icons, earcons, tactons and olfactory messages. This study grouped the modalities, focussing on the sensory channel that would receive the notification. The aim of the study was to identify if delivery to some sensory channels would be more disruptive or effective than to others, whether there were significant subjective workload differences between the modalities, and whether modal learning preference had an effect on the ability to receive and process messages in different modalities. The results of this study demonstrated that different modalities had very different properties, yet none created significantly more disruption or a higher workload. This work answered Thesis Question 1 by showing that, while visual and audio methods were generally the most effective, all the notification delivery methods tested performed well and could be used in a home care scenario.

Chapter 4 built on the results of the first study by considering each of the modalities separately and examining the relationship between performance and participant age. Disruption, effectiveness and subjective workload were examined but the experimental design was modified to allow another factor to be included: distraction. In addition, new measurements were developed to provide a deeper insight into the process of starting and stopping a task. The same modalities used in the first study were re-evaluated. The study showed that age had a very large impact on performance in almost every area. Much like the first study none of the modalities were demonstrably more disruptive or distracting, and workload did not appear to change significantly between the modalities. Chapter 4 answered Thesis Question 2 by providing useful data on the way different modalities delivered notifications.

Chapter 5 answered Thesis Question 3 by detailing the construction of a prototype dynamic multimodal reminder system called Dyna-Cue. To build Dyna-Cue, the results of the work carried out in Chapters 3 and 4 were combined with existing work. The work carried out in this chapter included the identification of the requirements for a dynamic prototype reminder system that can make dynamic decisions about *how* to interact in real-time based on the user, the environment and the configuration. Chapter 5 then went on to show how the implementation of Dyna-Cue met those requirements, exploring the internal architecture in detail.

Chapter 6 presented the final study of the thesis, which was designed to evaluate the Dyna-Cue prototype against alternative methods of reminder delivery. Specifically, the Dyna-Cue prototype was compared to a static reminder system (which always used text and speech) and a random system (which was equipped with multiple modalities, but randomly decided which one to use). A home-style variant of the Hotel Test by Manly *et al.* [133] was developed. This modified test was designed to stress prospective memory, responsible for managing and organising time and activities. Prospective memory is particularly sensitive to natural cognitive decline with age [188, 169]. The aim of the study was to identify which reminder delivery models were the most effective at aiding time management, resulted in the lowest subjective workload and were considered most appropriate by participants. Both younger and older participants were used and the effects of age were evaluated. The results showed that all three reminder technologies performed well for the younger participants, while the static technology was superior for the older participant group. The subjective feedback provided several important insights into the way participants perceived the technology, emphasising the need for personalisation and customisation. The Dyna-Cue prototype was discussed in light of the findings, generally focussing on how it could be improved and expanded on to deal with now situations. Given the ability of the Dyna-Cue system to provide multiple interaction models (*i.e.* it could be configured as a static system) it was concluded that dynamic multimodal models are the most appropriate design for use in a home care scenario, answering Thesis Question 4.

7.2 Contributions

The work presented in this thesis has made several contributions to the fields of multimodal interaction, notification systems and home care technology. This section will examine each of the thesis questions and discuss how the work presented addressed them. This section also includes a compilation of the guidelines defined in the thesis in Section 7.2.5. Section 7.2.6 lists additional contributions made by the thesis not directly linked to the thesis questions.

7.2.1 Thesis Question 1

Which forms of interaction are appropriate for use in a home care system?

This question was addressed by the work in Chapters 3 and 4, with additional insights being provided by the work carried out in Chapter 6. There are several forms of interaction that are appropriate for use in a home care system. Due to the differences between people and their homes, different interaction methods will be suitable in different situations. The work carried out into smell and tactile notification delivery methods in Chapters 3 and 4 showed that they were generally less effective than their visual and aural counterparts. However, they still performed quite well, and could be applied in situations where visual and aural notifications could not (*e.g.* when interacting with a deaf-blind person).

An interesting insight from the work in Chapter 4 was that for the visual and audio modalities, abstraction did not appear to result in poorer performance. This suggested that for simple messages abstract methods can be as effective as explicit ones for both younger and older people. This suggests that abstract modalities could be suitable for delivering sensitive messages that the user would not want others to understand.

In conclusion, there is a wide range of interaction methods suitable for use in a home care system, as the methods tested generally performed very well. As such, the appropriateness of an interaction method should be based on the unique needs of the user, not the performance of the interaction method.

7.2.2 Thesis Question 2

How do different forms of notification delivery affect users?

Answering this thesis question would provide data on how the user reacted to a notification delivered in different modalities. This was addressed by the work in Chapter 4, which expanded heavily on the work of Chapter 3 by providing new metrics and a more thorough analysis. There were several factors that might have affected the users, but it was modality that was of primary interest. The role of notification function (disruption *vs.* distraction) was also explored in Chapter 4.

The work carried out in Chapter 4 showed that in general there was little difference in the way participants reacted to notifications delivered in different modalities. Tactile notifications appeared to behave somewhat unusually, as there was evidence of a delay in responding to tactile messages. The results generally confirmed the findings of Chapter 3: that the modality of a notification does not appear to made a significant difference to disruption. Due to the more advanced metrics used in Chapter 4, it could also be concluded that the same applies for distractions.

Distractions were particularly interesting, because participants reacted more quickly to useful notifications (those requiring a response) than to useless distractor notifications. However, the processing times and long-term effects were the same for both types of notification. This suggests that participants become aware of useful information in the environment more quickly, but still suffer the same effects when faced with a distraction. The results highlight the importance of avoiding distractions in home care technology.

While this thesis question is very broad, the work carried out revealed important performance data for a range of modalities, allowing a conclusion to be drawn. In general the modality used to deliver a notification does not affect how disruptive or distracting that notification is. Chapter 4 also identified age as a significant factor, with older participants demonstrating much poorer performance compared to younger participants.

7.2.3 Thesis Question 3

> *How can home reminder technology be designed to best utilise multiple types of interaction?*

This thesis question was addressed in Chapter 5, which described how a piece of technology could automatically make use of any interaction methods it is equipped with, as long as those interaction methods provide a small amount of data to the reminder system. Chapter 5 outlined several of the core requirements that reminder technology should meet before describing the design, implementation and configuration of a prototype that met those standards.

In conclusion, the best way for a reminder system to utilise multiple types of interaction is to separate the message from the delivery method (a mode-independent representation [171, 174]), theoretically allowing a message to be delivered by any delivery method. It is possible to automatically combine delivery methods, and the properties of those combined methods can be calculated from their individual properties. This creates a wider range of options for notification delivery, all of which can potentially deliver the notification due to the mode-independent representation. Finally, knowledge of the properties of the different delivery methods can be used to make real-time decisions about which ones should be used when faced with different situations.

7.2.4 Thesis Question 4

> *Can home reminder technology be made more effective and appropriate by providing it with the ability to dynamically select from multiple forms of interaction?*

Chapter 6 answered this question by comparing the Dyna-Cue dynamic delivery method to static and random variants, along with a condition without any reminders at all. The results of that work showed that the Dyna-Cue system performed well for many users, but so did the static and random variants. Subjective feedback revealed a wide range of preferences, with few participants preferring to have no reminders at all. The results suggested that the dynamic behaviour of the Dyna-Cue system would not appeal to all users, with many preferring the static alternative.

This addressed Thesis Question 4 by showing that dynamic reminder technology can improve the effectiveness and appropriateness of home care technology. However, the Dyna-Cue prototype did not provided the required level of dynamism or customisation: the maximise effectiveness and appropriateness, the underlying logic of the system itself needed to be dynamic and customisable. Those options could only be offered by technology which includes multiple delivery methods, and would rely on some of the technology built into the Dyna-Cue prototype (such as the ability to automatically combine different delivery methods). In conclusion, dynamic multimodal reminder technology can be used to improved the effectiveness and appropriateness of home care technology.

7.2.5 Guidelines

Several guidelines were produced from the work carried out in Chapters 3, 4 and 6. This section presents a compilation of those guidelines.

- Drops in primary task performance post-notification appear to come from a lower rate of activity for certain modalities, and not an increased error rate post-notification.

- While modality does not appear to affect disruption, older people may be more susceptible to cross-modal interference, so care should be taken to avoid sensory channels being used by ongoing tasks.

- For simple messages, abstraction appears to have little effect (once training has taken place). Chapter 4 showed the abstract visual and audio modalities (*i.e.* abstract-visual, earcons) performed as well as the explicit modalities (*i.e.* text, speech). This can be used to deliver private information in a more discreet manner.

- Distractor notifications will cause negative effects that are similar to useful information, and as such their presence should be minimised as much as possible.

- Age has a significant effect on the ability to respond to notifications while carrying out an ongoing task. If notifications are being developed for older participants, the negative effects of both interruptions and distractions will be more severe. This will be more pronounced if the tactile and olfactory modalities are used.

- Assuming no significant impairment or background interference, visual and audio interaction methods provide the quickest and most effective ways to deliver information. Visual and audio modalities should be given preference for most notifications, unless there are special circumstances (*e.g.* sensory impairment).

- There is an additional subjective workload involved when using tactile and olfactory methods to deliver notifications.

- Tactile notification are less likely to be missed than olfactory notifications, but are more likely to be misunderstood.

- Tactile notifications are tricky; 'tactile lag' appears to exist and is manifest by a slightly slower delivery time, a longer processing time, and a difficulty in differentiating distractor notifications from target notifications. While this suggests that tactile notifications should not be used as a primary interaction method, the real-world effects of this 'tactile lag' are not significant enough to warrant avoiding use of the modality. Tactile should only be avoided if urgent notification delivery is required.

- Olfactory notifications are not highly salient and are likely to be missed. When noticed however, participants appeared to have little trouble understanding their meaning.

- Olfactory notifications should generally be avoided, as they produced the longest delivery times and the poorest response accuracy scores. However, performance was not so poor as to prevent olfactory notifications from being used in certain scenarios (*e.g.* to deal with severe impairment, or for non-urgent messages). Given their poor performance, olfactory notifications should not be used for important messages.

- Olfactory notifications appear to be more effective with female participants. However, the performance difference was not large enough that olfactory notifications should be avoided when delivering information to males.

- Well-timed reminders can be highly effective at helping to manage time and task management regardless of the modality used. This is particularly true for older participants.

- Understanding the properties of the modalities is key to making a dynamic multimodal reminder system. In Chapter 6, the tactile messages were frequently missed by older participants, which was considered to be responsible for the performance difference between the static and dynamic systems for older participants.

- A dynamic reminder system might make more intelligent scheduling decisions, but personal preference plays the pivotal role in determining appropriateness for the home and care. A dynamic system would need to provide significant customisation options to ensure that it could satisfy the user's needs.

- The results of Chapter 6 showed that the most appropriate methods of reminder delivery were the static and dynamic methods. As the dynamic technology includes the static and random alternatives, more technology should be dynamic to provide more options to the end-user. An important part of this is including multiple modalities.

7.2.6 Other Contributions

The work carried out in Chapters 3 and 4 contributed a new experimental design useful for experiments investigating the effects of interruptions and distractions. Performance data regarding eight modalities was gathered including average delivery speed and effectiveness, as well as a demonstration of how olfaction (a rarely explored communication method) can be used in practice. Chapter 3 also provided an exploration into the potential role of modal learning preference in determining the performance of multimodal notifications, although further work would be needed in this area.

Chapter 5 presented the Dyna-Cue system, a prototype multimodal reminder system and in itself another important contribution of this work. Chapter 5 explored the architecture of the Dyna-Cue system and included a justification of the rules the Dyna-Cue system is based on; those rules form another major contribution of this thesis, as they show how knowledge of a modality's properties can be combined with a simple set of rules to create useful dynamic behaviour.

Chapter 6 also included a reflection on the design of the Dyna-Cue prototype, identifying areas for further development. Another important contribution from this chapter was the development of a home-style prospective memory test that can be used to demonstrate the effectiveness of reminder technology in a controlled manner while maintaining ecological validity.

7.3 Limitations

There are several areas where the work carried out could have been improved. In Chapter 3, the different modalities were grouped by sensory apparatus instead of the individual modalities. One of the reasons this approach was taken was to reduce the complexity of the experiment, and it was also believed that the differences between sensory apparatuses would be quite similar, while the modalities would be similar enough to be grouped together. The findings in Chapter 4 suggest this is true, however it caused some problems interpreting the audio results. Due to the mixed-models way the visual and audio conditions were administered, there were too few participants within each modality for a between-groups analysis. This was rectified for the subsequent experiments in Chapter 4. While the design made *post hoc* analysis difficult, it did not affect the validity of the results.

In both Chapters 3 and 4 a large number of conditions were included, which complicated the process of carrying out *post hoc* pairwise comparisons (particularly in Chapter 4). However, suitable corrections were applied (*i.e.* Sidak and Tukey methods instead of Bonferroni where possible), and the inclusion of a control condition provided suitable context to confirm that the minor variations between experimental conditions were not type II errors. While the experimental designs used in Chapters 3 and 4 were complex, many conditions were necessary to create a comprehensive overview of the possibilities for multimodal interaction.

One shortcoming of the second study (Chapter 4) was the composition and size of the older user group. The older user group contained 16 people, yet 20 were desired for the experiment. There were issues finding enough participants for the older user group, which resulted in the data being analysed after 16 participants in order to meet a conference deadline. There were also two issues with the age of the older group: *(1)* the exact age of participants was not gathered, with an 'age group' being recorded instead and *(2)* the older group was aged 50 and over, which is at the lower end of the 'older user' age spectrum. The first point was only an issue because it became impossible to calculate the mean age of the older participant group. The second issue was raised when the age cut-off was criticised as being too low by conference paper reviewers. Despite this the results clearly showed a significant effect of age, with the older groups performance being almost half that of the younger group. While 50 and over was enough to demonstrate a significant age effect, the lower limit was raised to 60 for the work in Chapter 6.

In Chapter 6, the Dyna-Cue prototype did not perform as hoped during the evaluation. In Section 6.7 it was noted that this may have been due to older participants failing to notice tactile interactions. While the data appeared to support this hypothesis[1], more rigorous pilot studies may have prevented this from becoming an issue. Unfortunately, due to the difficulty in finding participants over 60 the pilot group consisted of only young and middle-aged people.[2]

In Chapters 4 and 6, mixed-models ANOVAs were used to create a model that included both the between-groups and within-groups variables. For the non-parametric data (*e.g.* NASA-TLX [79]) it would have been ideal to use a similar method. Unfortunately, there is no set standard on how to carry out such an analysis. A non-parametric mixed models test was attempted using the robust testing approach described by Wilcox [220]. As noted in Appendix D.3, the results were not deemed to be reliable which resulted in the non-parametric data being analysed with a range of tests instead of a single model.

One general shortcoming of the work in the thesis is that there is a lack of 'in the wild' research or longitudinal research in home. Due to the nature of the research, it was desirable to have tight controls and a large amount of reliable experimental data, which could not have been achieved to the same level in the home. There are several insights (particularly in Chapters 4 and 6) that would have been very difficult to observe in an 'in the wild' experimental design. However, real end-users participated in the studies and the work in Chapter 6 attempted to find a good balance between ecological validity and a controlled experiment. While it would have been useful to

[1] The data appeared to show a correlation between the drop in performance and the probability of receiving a tactile notification, but there is not enough data to confirm this.

[2] No data was gathered on pilot participants.

have carried out 'in the wild' studies alongside the work presented here, the lab-based designs used in the thesis were appropriate given the aims of the research and the type of data gathered. In addition, there would have been complex ethical issues if testing in the home with real end-users; performing lab-based studies is a reasonable compromise in this situation.

7.4 Future Work

This thesis presented a large amount of groundwork research on the creation of a dynamic multimodal reminder system. Yet there is a great deal of further work that needs to be carried out to further develop the concept of dynamic multimodal reminder technology. Future work based on this thesis should include:

- Further work introducing new modalities. The methodology detailed in Chapter 4 could be repeated for individual modalities, allowing new data to be gathered for modalities that were not part of the original experiment. Thermal interaction [223] is an example of a new modality of interest, as it is discreet and different temperatures could be used to provide an interesting range of salience.

- Further development on the Dyna-Cue prototype as discussed in Section 6.9, in particular:

    - implementing OSGI so that modalities can be added and removed while the system is running;
    - implementing a functional CI (most probably using the Microsoft Kinect) and using more fine-grained data to adjust the properties of the delivery methods in real-time (*e.g.* by changing salience scores based on attentional focus);
    - include user preferences in the technology, in particular implementing ways to define the modality selection logic, *i.e.* allowing the user to choose between dynamic modality selection and static modality selection.

- Longitudinal studies of the Dyna-Cue prototype in the home with real end-users.

- Testing of the Dyna-Cue prototype with impaired users, specifically users with sensory impairments.

- Research ways to allow configuration of the system's *behaviour* by real-world end-users, *i.e.* which modalities it would use and when.

- Research ways to allow configuration of the system's *schedule* by real-world end-users, *i.e.* how the user can set the messages to deliver and the message triggering conditions. One way that was considered for the configuration of the system was a smart pen which could upload data from a pen-and-paper calendar [221].

One shortcoming of this work is that it was carried out under controlled conditions. Longitudinal studies in the home would be a logical next step to further explore this technology. For a home-based study the the primary areas of interest would be how effective the system was at helping the user, and how the user felt about having the technology in their home. While different stakeholders are likely to have different opinions on effectiveness (as discussed in Section 2.4.2), for an initial study it the most appropriate measure would be schedule compliance. The best way to measure this would be either by recording the user's reactions to the system or by setting up appropriate sensors to log compliance. Self-reporting of compliance would not be appropriate for a home-trial; it would immediately introduce a complex confounding factor (as it would impact on home routines) and the data is likely to be unreliable. As shown by the conclusions of Chapter 6, it's important to gather multiple forms of subjective data. While Likert scales and similar measures are useful for statical analyses, a face-to-face interview is vital for understanding the reasoning behind the participant's results. While studies like this lack the tight controls of a laboratory setting, they are vital to demonstrate that the the lab-based findings can indeed be transferred to the home environment and should be considered the most interesting area for future work.

Outside the work of Vastenburg *et al.* [203, 202] and Perry *et al.* [171], there is very little existing work on dynamic multimodal technology for the home. This thesis presents an in-depth analysis of a range of modalities along with an evaluated reminder system prototype. As this research area is still relatively young, it is hoped that the work presented will form an important contribution to the groundwork needed for the creation of fully-functional dynamic multimodal reminder technology for home care.

7.5 Conclusions

It was hypothesised at the start of this thesis that multiple modalities would help to overcome sensory impairment and make technology for home care more accessible and robust. More options for interaction would allow for more opportunities for customisation, which is vital to making any technology for the home acceptable to its users. Perhaps most importantly, multiple modalities would provide the level of built-in flexibility needed to compensate for the rapidly changing requirements that are often found in a home care setting.

The Dyna-Cue prototype has demonstrated the validity of these theories. The static and random variants of the Dyna-Cue prototype were preferred by some participants, while others preferred the Dyna-Cue dynamic modality selection model. While the Dyna-Cue prototype is capable of delivering all three models, the inverse is not also true. As such the Dyna-Cue prototype demonstrates the level of flexibility that dynamic multimodal technology should offer. It has also demonstrated the power of changing *how* interactions take place: by switching from an often ineffective tactile message to other modalities, the Dyna-Cue prototype was able to ensure good performance.

The Dyna-Cue system and its evaluation are one of the main contributions of this thesis. However, the Dyna-Cue system could not have been constructed without a thorough understanding of the properties of a range of different modalities, which was provided by the work in Chapters 3 and 4. This work has made an important contribution to the fields of multimodal interaction, interruption management and home care technology.

The results of this work will help to address the economic and ethical issues raised by a growing older population. This thesis has demonstrated the benefits that multimodality and intelligent scheduling can bring. As such, home care technology must aspire to include as many modalities as possible to maximise the effectiveness of the technology and its appropriateness for the home.

Appendix A

Additional Materials – Chapter 3

This appendix provides additional materials related to the first study of this thesis, which is described in Chapter 3. Source code, raw data and analysis data are provided in the accompanying materials as described in Appendix E.

A.1 Experiment Materials

A.2 Notification Response Accuracy Full Results

Section 3.6.5 assessed the accuracy of responses to the notifications in different modalities. This section of the appendix provides additional graphs showing the results and tables including pairwise comparisons of the conditions.

A.3 NASA-TLX Full Results

Section 3.6 presented the results of the NASA-TLX. The full results of the NASA-TLX were omitted for brevity. Full results are presented here. See also Figure 3.19 on page 104, which may clarify how the overall workload is calculated from the component parts.

The MultiMemoHome Project

Multimodal Reminders in the Home

You are invited to take part in a research study to explore different types of electronic reminders. You will be asked to play a simple card matching game (‘Concentration’ on a computer and while you are playing this game you might receive a ‘reminder’ message with instructions on what action to perform. The aim of the game is to turn over two cards at a time in order to find ‘matching pairs’. When a pair *is not* matched on any one turn, the cards will remain turned down until you select them again on any other turn(s). When a pair *is* matched correctly in any one turn, the cards will remain blanked out. The aim of the game is to find all the pairs as quickly as you can with as few turns of the cards as possible.

The reminder messages, when presented, will arrive in different ways - visually (a text message on the screen, an icon, or a colour on the screen), aurally (speech, an earcon (ring tone sound), or an auditory icon (sound effect)), tactile (a vibration to the skin or a temperature message to the skin), olfactory (a smell or aroma).

When you receive a message during game play you should press the button that corresponds to the action you were instructed to carry out. You should attend to the message as quickly and as accurately as you can; you will have a limited amount of time to repond.

We are interested in finding out about the different types of reminders and how people play the game. You do not need to have technical expertise to take part. You can be any age to take part but we are particularly keen to include people over 50.

Your participation in this study will be confidential. All tapes and notes will be put in locked filing cabinets: computer data will be stored securely under passwords in compliance with the data protection protocols. You will not be identified personally in reports or publications unless we have asked you for your permission for this directly.

We hope your participation will help us improve the design of computer technology for the home. We also aim to improve the process of providing reminding technology solutions. We wish to ensure these are appropriate and usable. We also wish to make sure these are suited to each person's needs, living circumstances, and preferences. We hope that you will enjoy learning about new research ideas and technologies.

The research will mainly be carried out by the following people:

- Mr David Warnock, a PhD student at the University of Glasgow
- Dr. Marilyn McGee-Lennon, a researcher at the University of Glasgow

Your participation is voluntary. You can ask us questions about the study before you decide whether to participate. You can withdraw from the study at any time without any problems. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

To arrange to take part in the study, or if you have any questions, please contact:

David Warnock

Computing Science
University of Glasgow
8–17 Lilybank Gardens
Glasgow G12 8QQ

Email: warnockd@dcs.gla.ac.uk

You can also visit the MultiMemoHome website for more information on what we do:

http://www.multimemohome.co.uk

MMH The MultiMemoHome Project

Multimodal Reminders in the Home

Please initial box

1. I confirm that I have read and understand the information sheet provided. I have had the opportunity to consider the information and to ask questions. I have had these answered satisfactorily.

2. I understand that my participation is voluntary, and that I am free to withdraw at any time without giving any reason, I understand that my care or legal rights will not be affected.

3. I understand that the information collected in this study will be confidential, anonymous and protected.

4. I understand that taking part in this study will not have a directly measurable effect on my health or well-being.

5. I understand that my responses will be recorded but anonymised. I give permission for this information, including the use of quotations, to be used in any presentation of the research. I understand that my anonymity will be assured.

6. I agree to take part in the above study

-------------------------------------	----------------	---------------------------
Name of participant	Date	Signature
-------------------------------------	----------------	---------------------------
Name of research team present	Date	Signature

Figure 8.6

NASA Task Load Index

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

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Sensory Modality Preference Inventory

Read each statement and select the appropriate number response as it applies to you.

Often (3)	Sometimes (2)	Seldom/Never (1)

Visual Modality

_____ I remember information better if I write it down.
_____ Looking at the person helps keep me focused.
_____ I need a quiet place to get my work done.
_____ When I take a test, I can see the textbook page in my head.
_____ I need to write down directions, not just take them verbally.
_____ Music or background noise distracts my attention from the task at hand.
_____ I don't always get the meaning of a joke.
_____ I doodle and draw pictures on the margins of my notebook pages.
_____ I have trouble following lectures.
_____ I react very strongly to colors.

_____ **Total**

Auditory Modality

_____ My papers and notebooks always seem messy.
_____ When I read, I need to use my index finger to track my place on the line.
_____ I do not follow written directions well.
_____ If I hear something, I will remember it.
_____ Writing has always been difficult for me.
_____ I often misread words from the text, i.e. "them" for "then").
_____ I would rather listen and learn than read and learn.
_____ I'm not very good at interpreting an individual's body language.
_____ Pages with small print or poor quality copies are difficult for me to read.
_____ My eyes tire quickly, even though my vision check-up is always fine.

_____ **Total**

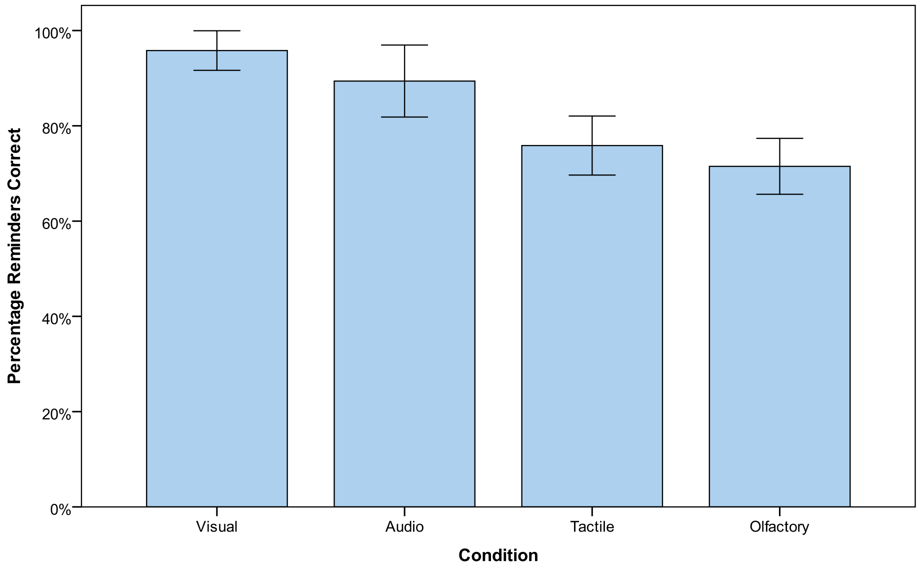
Kinesthetic/Tactile Modality

_____ I start a project before reading the directions.
_____ I hate to sit at a desk for long periods of time.
_____ I prefer first to see something done and then to do it myself.
_____ I use the trial and error approach to problem-solving.
_____ I like to read my textbook while riding an exercise bike.
_____ I take frequent study breaks.
_____ I have a difficult time giving step-by-step instructions.
_____ I enjoy sports and do well at several different types of sports.
_____ I use my hands when describing things.
_____ I have to rewrite or type my class notes to reinforce the material.

_____ **Total**

Total the score for each section. A score of 21 points or more in a modality indicates a strength in that area. The highest of the 3 scores indicates the most efficient method of information intake. The second highest score indicates the modality that boosts the primary strength. For example, a score of 23 in the visual modality indicates a strong visual learner. Such a learner benefits from the text, from filmstrips, charts, graphs, etc. If the second highest score is auditory, then the individual would benefit from audio tapes, lectures, etc. If you are strong kinesthetically, then taking notes and rewriting class notes will reinforce information.

Figure A.1: Additional data on the percentage of notifications correctly responded to.



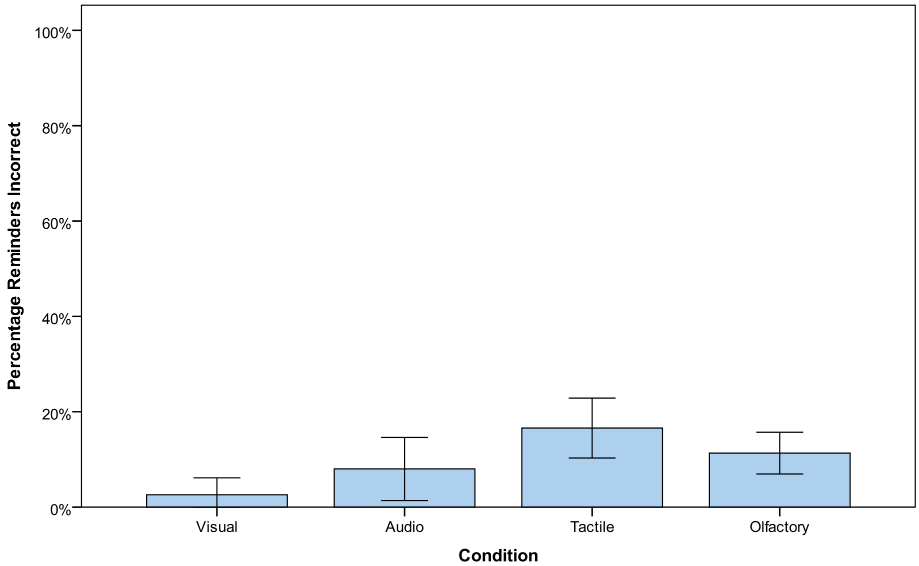
(a) Graph showing the relationship between sensory channel and the percentage of notifications correctly acknowledged. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons of the effect of sensory channel used on the percentage of notifications correctly acknowledged.

	Vis		Aud		Tac		Olf	
Vis	-		.588		.000	***	.000	***
Aud	.588		-		.246		.065	
Tac	.000	***	.246		-		.986	
Olf	.000	***	.065		.986		-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.2: Additional data on the percentage of notifications incorrectly responded to.



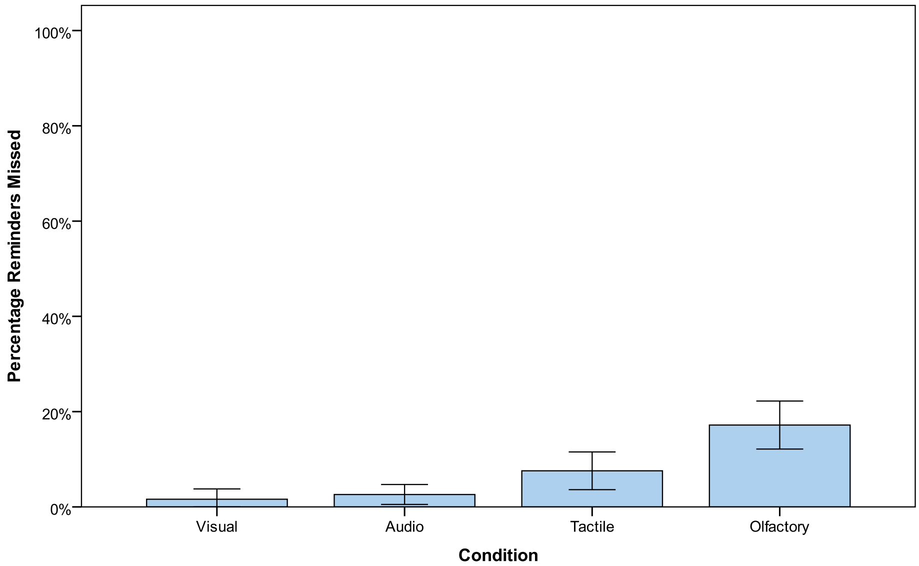
(a) Graph showing the relationship between sensory channel and the percentage of notifications incorrectly acknowledged. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons of the effect of sensory channel used on the percentage of notifications incorrectly acknowledged.

	Vis		Aud	Tac		Olf	
Vis	-		.588	.009	**	.004	**
Aud	.588		-	.691		.998	
Tac	.009	**	.691	-		.661	
Olf	.004	**	.998	.661		-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.3: Additional data on the percentage of notifications not responded to.



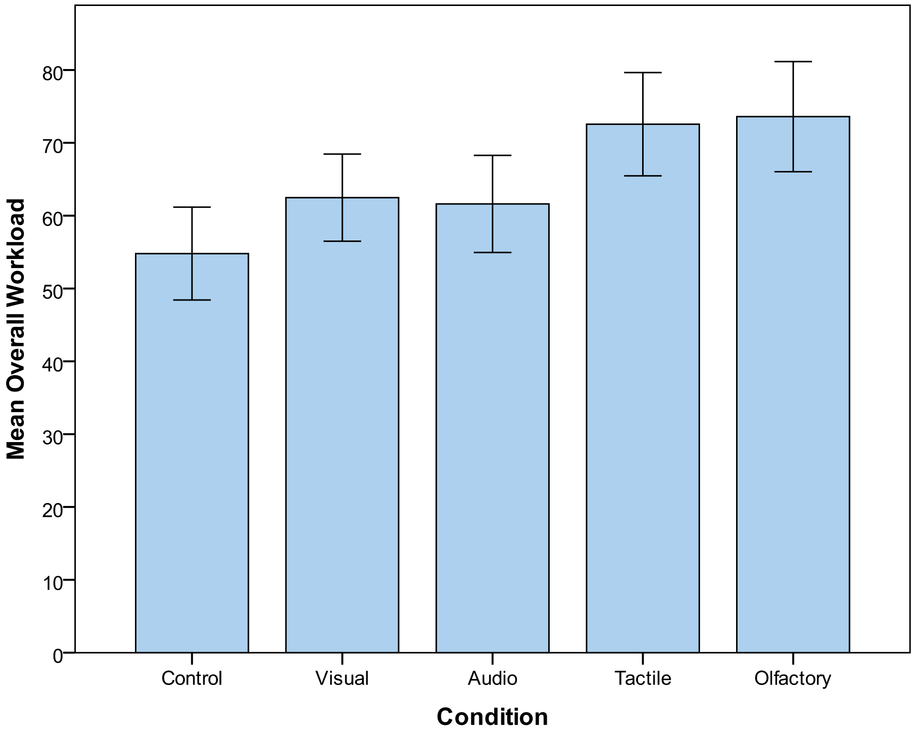
(a) Graph showing the relationship between sensory channel and the percentage of notifications unacknowledged. Error bars show 95% confidence intervals.

(b) Significance (p) values *post hoc* pairwise comparisons of the effect of sensory channel used on the percentage of notifications unacknowledged.

	Vis		Aud		Tac	Olf	
Vis	-		.979		.104	.000	***
Aud	.979		-		.184	.001	***
Tac	.104		.184		-	.222	
Olf	.000	***	.001	***	.222	-	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.4: The effects of the experimental condition on the overall workload associated with the task.



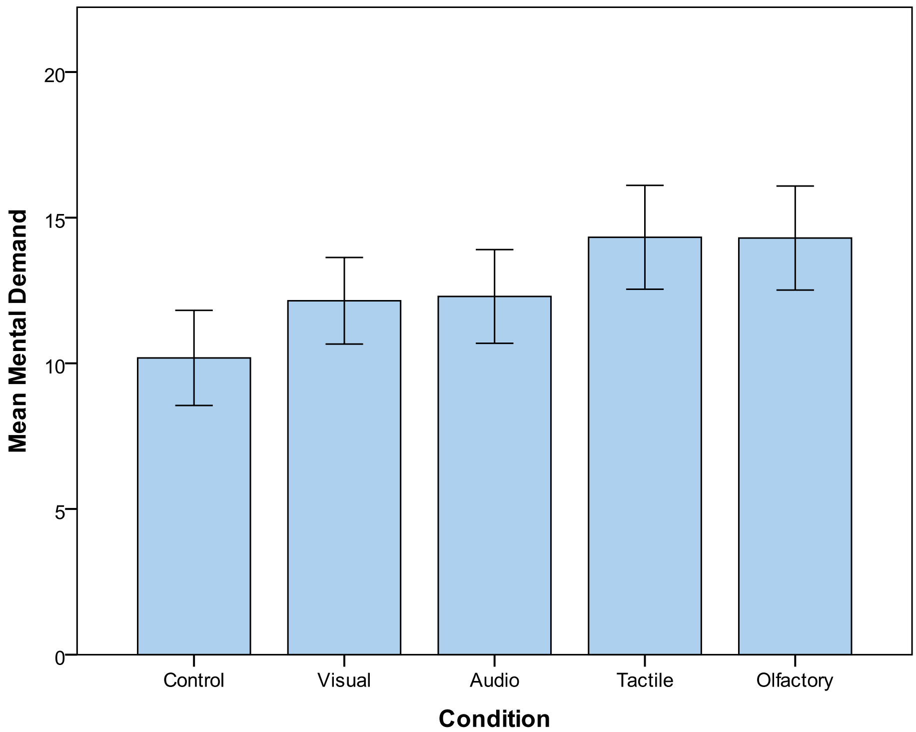
(a) Overall Workload. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the overall workload.

	Con		Vis		Aud		Tac		Olf	
Con	-		-.92		.90		-2.40	***	-2.35	***
Vis	-.92		-		-.02		1.48	*	1.44	*
Aud	.90		-.02		-		-1.50	*	-1.46	*
Tac	-2.40	***	1.48	*	-1.50	*	-		-.04	
Olf	-2.35	***	1.44	*	-1.46	*	-.04		-	

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.5: The effects of the experimental condition on the mental demand associated with the task.



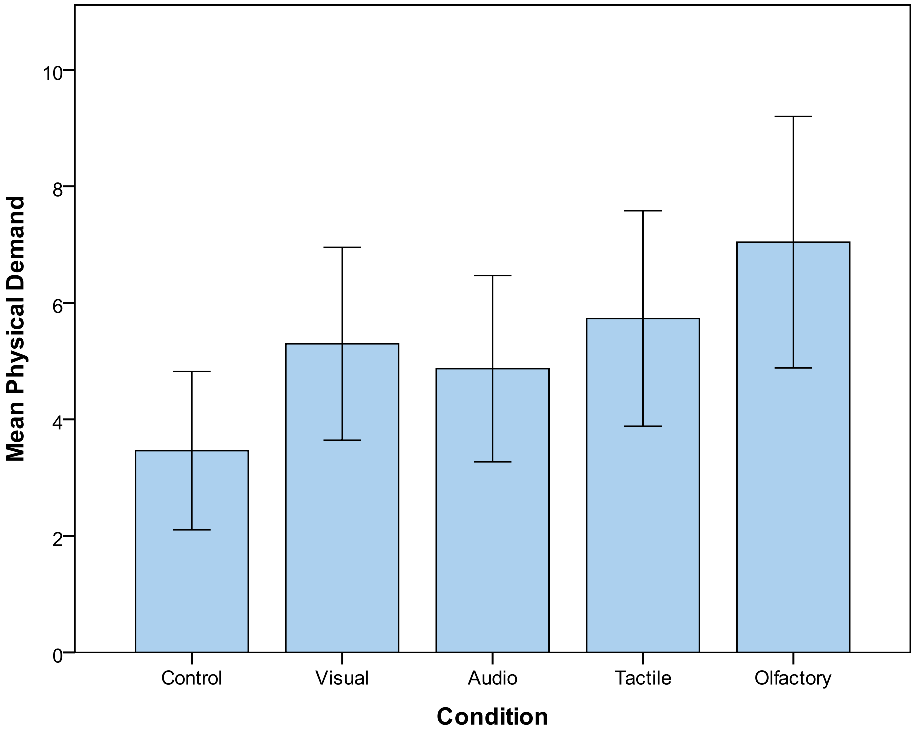
(a) Mental Demand. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the mental demand.

	Con		Vis		Aud		Tac		Olf	
Con	-		-1.90		1.33	*	-2.54	***	-2.54	***
Vis	-1.90		-		.15		1.35	*	1.35	*
Aud	1.33	*	.15		-		-1.21		-1.21	
Tac	-2.54	***	1.35	*	-1.21		-		.00	
Olf	-2.54	***	1.35	*	-1.21		.00		-	

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.6: The effects of the experimental condition on the physical demand associated with the task.



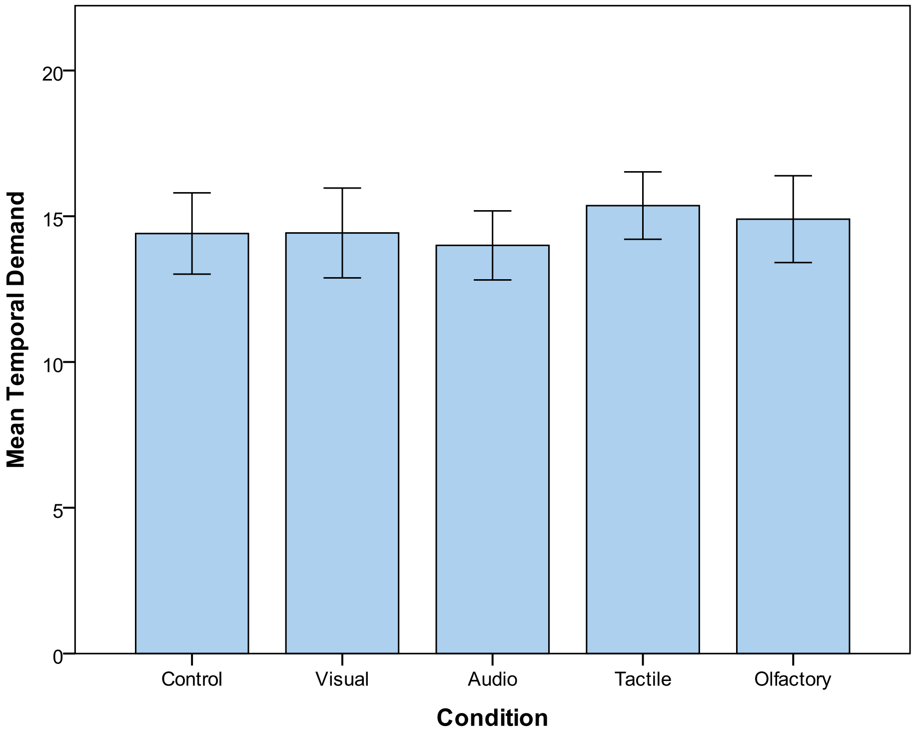
(a) Physical Demand. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the physical demand.

	Con		Vis	Aud	Tac		Olf	
Con	-		-1.10	1.00	-1.65	**	-2.08	***
Vis	-1.10		-	-1.04	.54		.98	
Aud	1.00		-1.04	-	-.65		-1.08	
Tac	-1.65	**	.54	-.65	-		.44	
Olf	-2.08	***	.98	-1.08	.44		-	

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.7: The effects of the experimental condition on the temporal demand associated with the task.



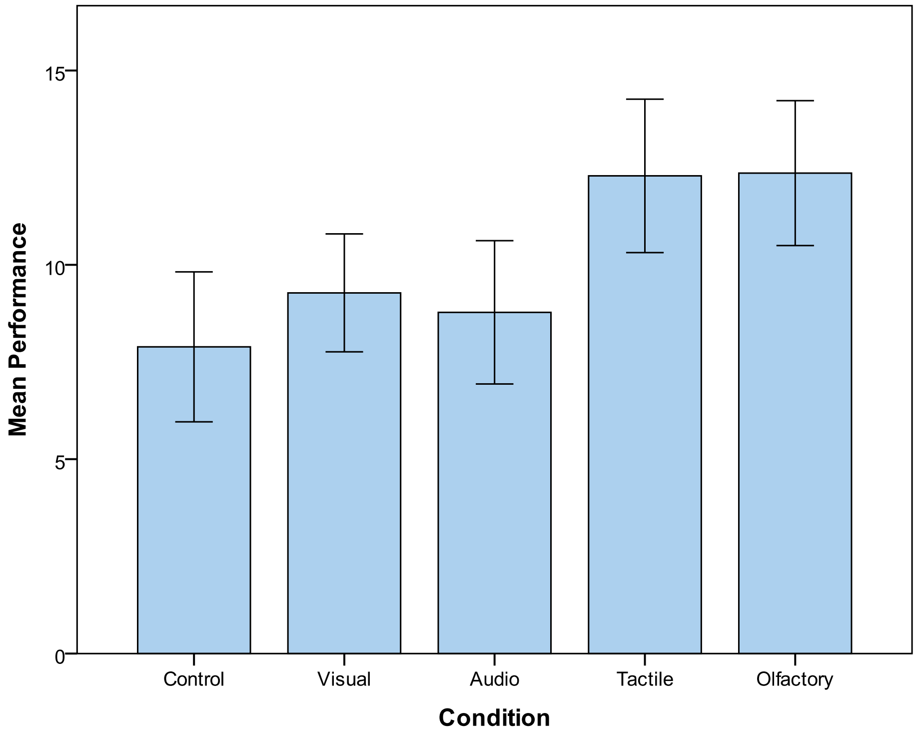
(a) Temporal Demand. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the temporal demand.

	Con	Vis	Aud	Tac	Olf
Con	-	.04	-.38	-.65	-.71
Vis	.04	-	-.33	.69	.75
Aud	-.38	-.33	-	-1.02	-1.08
Tac	-.65	.69	-1.02	-	.06
Olf	-.71	.75	-1.08	.06	-

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.8: The effects of the experimental condition on the performance level associated with the task.



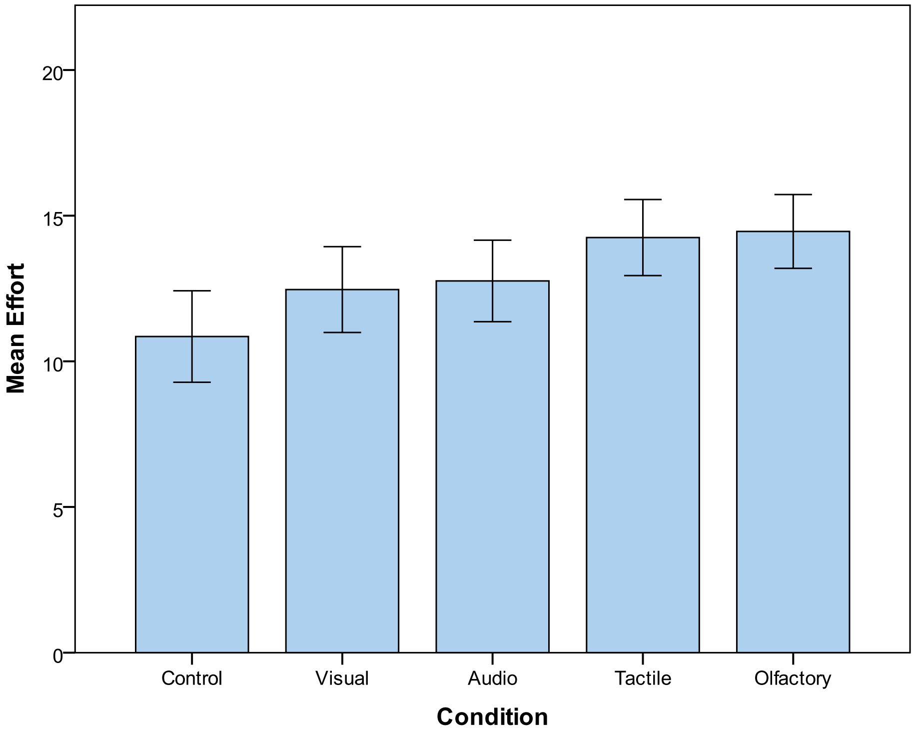
(a) Performance. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the performance level.

	Con		Vis	Aud		Tac		Olf	
Con	-		-1.02	.73		-2.29	***	-2.21	***
Vis	-1.02		-	-.29		1.27		1.19	
Aud	.73		-.29	-		-1.56	**	-1.48	*
Tac	-2.29	***	1.27	-1.56	**	-		-.08	
Olf	-2.21	***	1.19	-1.48	*	-.08		-	

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.9: The effects of the experimental condition on the effort associated with the task.



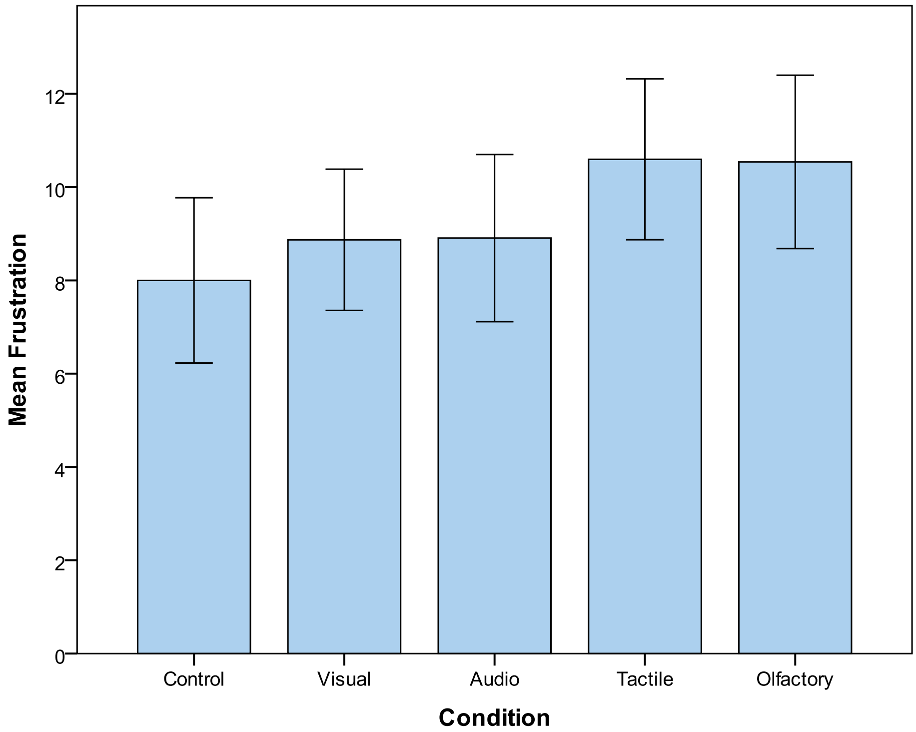
(a) Effort. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the effort required.

	Con		Vis		Aud		Tac		Olf	
Con	-		-1.08		1.19		-2.44	***	-2.48	***
Vis	-1.08		-		.10		1.35	*	1.40	*
Aud	1.19		.10		-		-1.25		1.40	*
Tac	-2.44	***	1.35	*	-1.25		-		.04	
Olf	-2.48	***	1.40		1.40	*	.04		-	

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure A.10: The effects of the experimental condition on the frustration level associated with the task.



(a) Frustration. Error bars show 95% confidence intervals.

(b) *Post hoc* pairwise comparisons showing the effects of the condition on the frustration level reported.

	Con		Vis	Aud	Tac		Olf	
Con	-		-.77	.56	-1.48	*	-1.35	*
Vis	-.77		-	-.21	.71		.58	
Aud	.56		-.21	-	-.92		-.79	
Tac	-1.48	*	.71	-.92	-		-.13	
Olf	-1.35	*	.58	-.79	-.13		-	

Note: Each entry shows the Friedman's ANOVA χ^2 value. Significance (p) values corrected using the Bonferroni method. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Appendix B

Additional Materials – Chapter 4

This appendix presents a full description of the materials used in the second study of this thesis, which is described in Chapter 4. Source code, raw data and analysis data are provided in the accompanying materials as described in Appendix E.

B.1 Experiment Materials

The NASA Task-Load Index (NASA-TLX) form used was identical to the one used in the first study, which is provided in Appendix A (p. 271). The consent form used was also identical (p. 270).

B.2 Disruption and Effectiveness: Additional Results

This section provides additional graphs and tables that were omitted from Section 4.6 for brevity.

B.3 NASA-TLX Full Results

Section 4.6 presented the results of the NASA-TLX. The full results of the NASA-TLX were omitted for brevity. Full results are presented here. See also Figure 4.13 on page 154, which may clarify how the overall workload is calculated from the component parts.

The MultiMemoHome Project

Multimodal Reminders in the Home

You are invited to take part in a research study to explore different types of computer-controlled notifications. You will be asked to play a simple card matching game (Concentration) on a computer while responding to certain notifications by pressing a button.

The aim of the game is to turn over two cards per turn in order to find matching pairs. When a pair *is not* matched on any one turn, the cards are turned back over until you select them again on any other turn(s). When a pair *is* matched correctly, the cards will be removed from play. The aim of the game is to find all the pairs as quickly as you can in as few turns as possible.

When presented the notifications will arrive in different ways; visually as a textual message, icon or colour; aurally as speech, earcons (ring-tone style sound) or auditory icons (sound effect); tactile (a vibration to the skin); or olfactory (a smell or aroma).

For each type of notification, only one will require a response; you will be shown which notification to respond to at the start. All other notifications should be ignored. Try to attend to the appropriate notifications as quickly and accurately as you can. To respond to a notification press the yellow button.

We are interested in finding out about the different types of notifications and how they affect the way people play the game. We have already examined the performance of younger participants, so we are now looking for participants who are over 50 years old.

Your participation in this study will be confidential. All notes will be kept in locked filing cabinets and all computer data will be stored securely under passwords in compliance with data protection protocols. You will not be identified personally in any reports or publications unless we have directly asked for your permission.

We hope your participation will help us improve the design of computer technology for the home, and that you will enjoy learning about new research ideas and technologies.

The research will mainly be carried out by the following people:

- Mr David Warnock, a PhD student at the University of Glasgow
- Dr. Marilyn McGee-Lennon, a researcher at the University of Glasgow

Your participation is voluntary. You can ask us questions about the study before you decide whether to participate. You can withdraw from the study at any time without consequence. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

To arrange to take part in the study, or if you have any questions, please contact:

David Warnock

Computing Science,
University of Glasgow,
8–17 Lilybank Gardens,
Glasgow, G12 8QQ

Email: **warnockd@dcs.gla.ac.uk**

Phone: **+44-141-330-4256 (ask for ext. 0672)**

You can also visit the MultiMemoHome website for more information on what we do.

http://MultiMemoHome.org

Table B.1: Significance (p) values of *post hoc* pairwise comparisons of the modality on the mean number of cards matched.

	Con	Text	Pict	Abs-Vis	Speech	Aud-Icon	Earcon	Tac	Olf
Con	-	.000 ***	.051	.004 **	.003 **	.041 *	.025 *	.002 **	.001 ***
Text	.000 ***	-	1.000	1.000	.909	.998	1.000	1.000	1.000
Pict	.051	1.000	-	1.000	1.000	1.000	1.000	.998	1.000
Abs-Vis	.004 **	1.000	1.000	-	1.000	1.000	1.000	1.000	1.000
Speech	.003 **	.909	1.000	1.000	-	1.000	1.000	.893	.999
Aud-Icon	.041 *	.998	1.000	1.000	1.000	-	1.000	.986	1.000
Earcon	.025 *	1.000	1.000	1.000	1.000	1.000	-	1.000	1.000
Tac	.002 **	1.000	.998	1.000	0.893	0.986	1.000	-	1.000
Olf	.001 ***	1.000	1.000	1.000	0.999	1.000	1.000	1.000	-

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Table B.2: Significance (p) values of *post hoc* pairwise comparisons of the modality on the error rate (superfluous views per turn).

	Con	Text	Pict	Abs-Vis	Speech	Aud-Icon	Earcon	Tac	Olf
Con	-	.000 ***	.008 **	.034 *	.536	.002 **	.104	.084	.001 ***
Text	.000 ***	-	.998	.986	.454	1.000	.514	.851	.999
Pict	.008 **	.998	-	1.000	.940	1.000	1.000	1.000	1.000
Abs-Vis	.034 *	.986	1.000	-	1.000	1.000	1.000	1.000	1.000
Speech	.536	.454	.940	1.000	-	.712	1.000	1.000	.917
Aud-Icon	.002 **	1.000	1.000	1.000	0.712	-	.997	.999	1.000
Earcon	.104	.514	1.000	1.000	1.000	0.997	-	1.000	.993
Tac	.084	.851	1.000	1.000	1.000	0.999	1.000	-	1.000
Olf	.001 ***	.999	1.000	1.000	0.917	1.000	.993	1.000	-

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Table B.3: Significance (p) values of *post hoc* pairwise comparisons on the accuracy of responses to notifications in different modalities.

	Text	Pict	Abs-Vis	Speech	Aud-Icon	Earcon	Tac	Olf
Text		1.000	.999	.734	.384	1.000	.268	.001 ***
Pict	1.000		1.000	.681	.089	1.000	.012 *	.000 ***
Abs-Vis	.999	1.000		1.000	1.000	.981	.020 *	.000 ***
Speech	.734	.681	1.000		1.000	.689	.001 **	.000 ***
Aud-Icon	.384	.089	1.000	1.000		.697	.000 ***	.000 ***
Earcon	1.000	1.000	.981	0.689	0.697		.846	.034 *
Tac	.268	.012 *	.020 *	0.001 **	0.000 ***	.846		.894
Olf	.001 ***	.000 ***	.000 ***	0.000 ***	0.000 ***	.034 *	.894	

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Table B.4: Significance (p) values of *post hoc* pairwise comparisons on notification processing time.

	Text	Pict	Abs-Vis	Speech	Aud-Icon	Earcon	Tac	Olf
Text		1.000	1.000	1.000	1.000	.799	.000 ***	.001 **
Pict	1.000		.999	.994	1.000	.815	.000 ***	.000 ***
Abs-Vis	1.000	.999		1.000	.995	1.000	.000 ***	.001 ***
Speech	1.000	.994	1.000		.998	.998	.000 ***	.003 **
Aud-Icon	1.000	1.000	.995	0.998		.806	.000 ***	.000 ***
Earcon	.799	.815	1.000	0.998	0.806		.003 **	.045 *
Tac	.000 ***	.000 ***	.000 ***	0.000 ***	0.000 ***	.003 **		.906
Olf	.001 **	.000 ***	.001 ***	0.003 **	0.000 ***	.045 *	.906	

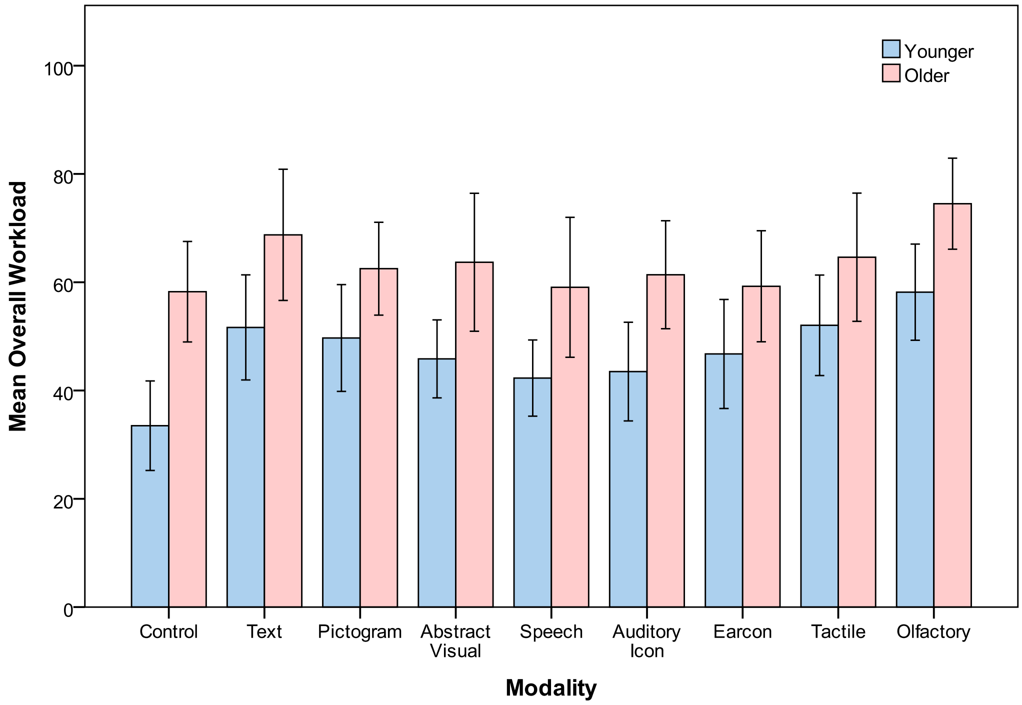
Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Table B.5: Significance (p) values of *post hoc* pairwise comparisons on notification delivery time.

	Text		Pict		Abs-Vis		Speech		Aud-Icon		Earcon		Tac		Olf	
Text			.019	*	.889		.775		.927		1.000		.001	**	.000	***
Pict	.019	*			.712		.280		.196		.023	*	.000	***	.000	***
Abs-Vis	.889		.712				1.000		1.000		.782		.000	***	.000	***
Speech	.775		.280		1.000				1.000		.830		.000	***	.000	***
Aud-Icon	.927		.196		1.000		1.000				.741		.000	***	.000	***
Earcon	1.000		.023	*	.782		0.830		0.741				.000	***	.000	***
Tac	.001	**	.000	***	.000	***	0.000	***	0.000	***	.000	***			.000	***
Olf	.000	***	.000	***	.000	***	0.000	***	0.000	***	.000	***	.000	***		

Note: Significance values calculated by SPSS using the estimated marginal means and corrected using the Sidak correction. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.1: Additional data for TLX overall workload.



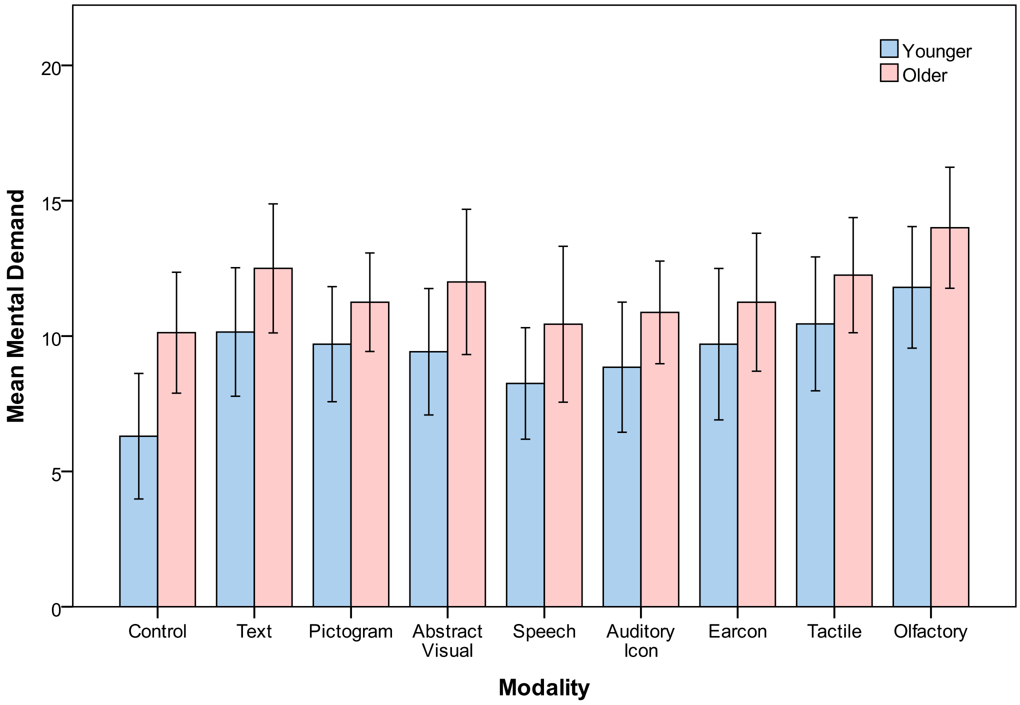
(a) The relationship between modality and overall workload. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the overall workload scores for the modalities.

Group		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.001***							
	Pic	0.004**	1.000						
	Abs	0.131	0.840	0.975					
	Spe	0.406	0.485	0.777	1.000				
	Aud	0.334	0.567	0.840	1.000	1.000			
	Ear	0.030*	0.985	1.000	1.000	0.978	0.990		
	Tac	0.000***	1.000	1.000	0.778	0.406	0.485	0.970	
	Olf	0.000***	0.931	0.725	0.105	0.021*	0.030*	0.333	0.960
Older	Tex	0.044*							
	Pic	0.978	0.490						
	Abs	0.693	0.912	0.999					
	Spe	1.000	0.118	0.999	0.888				
	Aud	0.988	0.423	1.000	0.997	1.000			
	Ear	1.000	0.139	0.999	0.912	1.000	1.000		
	Tac	0.558	0.963	0.993	1.000	0.793	0.985	0.828	
	Olf	0.019*	1.000	0.321	0.793	0.058	0.266	0.071	0.888

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.2: Additional data for TLX mental demand.



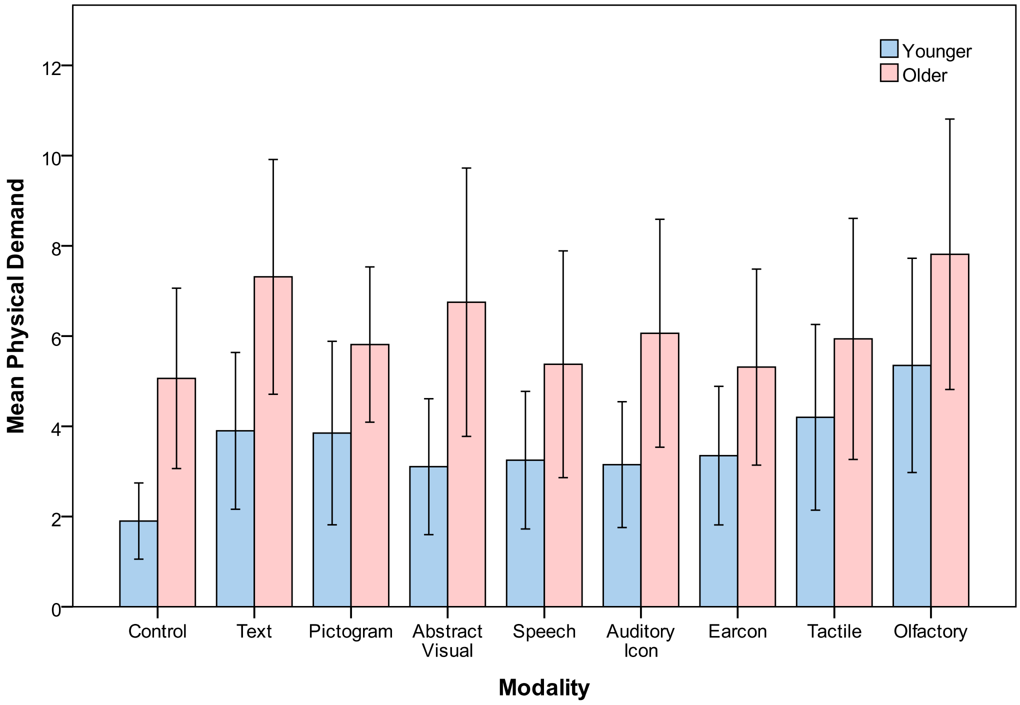
(a) The relationship between modality and mental demand. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the mental demand scores for the modalities.

Group		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.003**							
	Pic	0.022*	1.000						
	Abs	0.104	0.973	1.000					
	Spe	0.550	0.570	0.907	0.995				
	Aud	0.408	0.711	0.963	0.999	1.000			
	Ear	0.111	0.968	1.000	1.000	0.996	1.000		
	Tac	0.003**	1.000	1.000	0.973	0.571	0.711	0.968	
	Olf	0.000***	0.984	0.800	0.447	0.070	0.121	0.428	0.984
Older	Tex	0.177							
	Pic	0.958	0.888						
	Abs	0.480	1.000	0.993					
	Spe	1.000	0.503	0.999	0.843				
	Aud	0.989	0.772	1.000	0.970	1.000			
	Ear	0.827	0.980	1.000	1.000	0.986	1.000		
	Tac	0.110	1.000	0.791	0.998	0.370	0.643	0.943	
	Olf	0.006**	0.975	0.221	0.791	0.043*	0.130	0.435	0.993

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.3: Additional data for TLX physical demand.



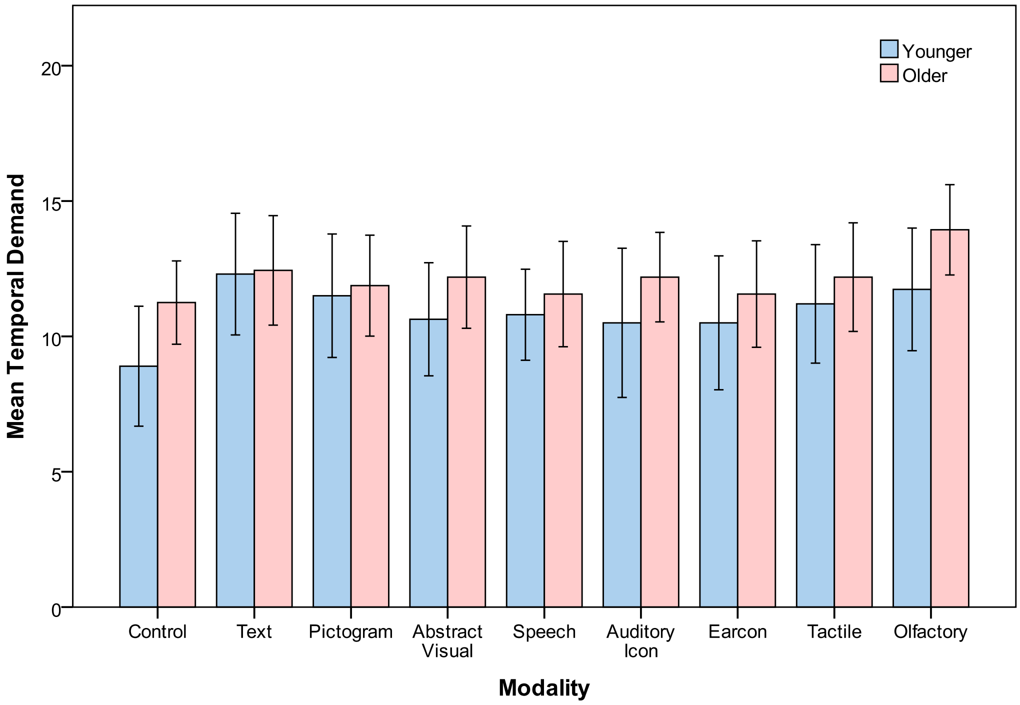
(a) The relationship between modality and physical demand. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the physical demand scores for the modalities.

Group		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.052							
	Pic	0.325	0.998						
	Abs	0.431	0.991	1.000					
	Spe	0.749	0.900	0.999	1.000				
	Aud	0.454	0.989	1.000	1.000	1.000			
	Ear	0.204	1.000	1.000	1.000	0.995	1.000		
	Tac	0.031*	1.000	0.991	0.975	0.825	0.970	0.999	
	Olf	0.001***	0.950	0.547	0.431	0.175	0.408	0.707	0.979
Older	Tex	0.017*							
	Pic	0.855	0.581						
	Abs	0.270	0.984	0.992					
	Spe	1.000	0.008**	0.742	0.174				
	Aud	0.987	0.252	1.000	0.885	0.958			
	Ear	1.000	0.021*	0.885	0.308	1.000	0.992		
	Tac	0.605	0.838	1.000	1.000	0.461	0.992	0.652	
	Olf	0.437	0.932	0.999	1.000	0.308	0.964	0.484	1.000

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.4: Additional data for TLX temporal demand.



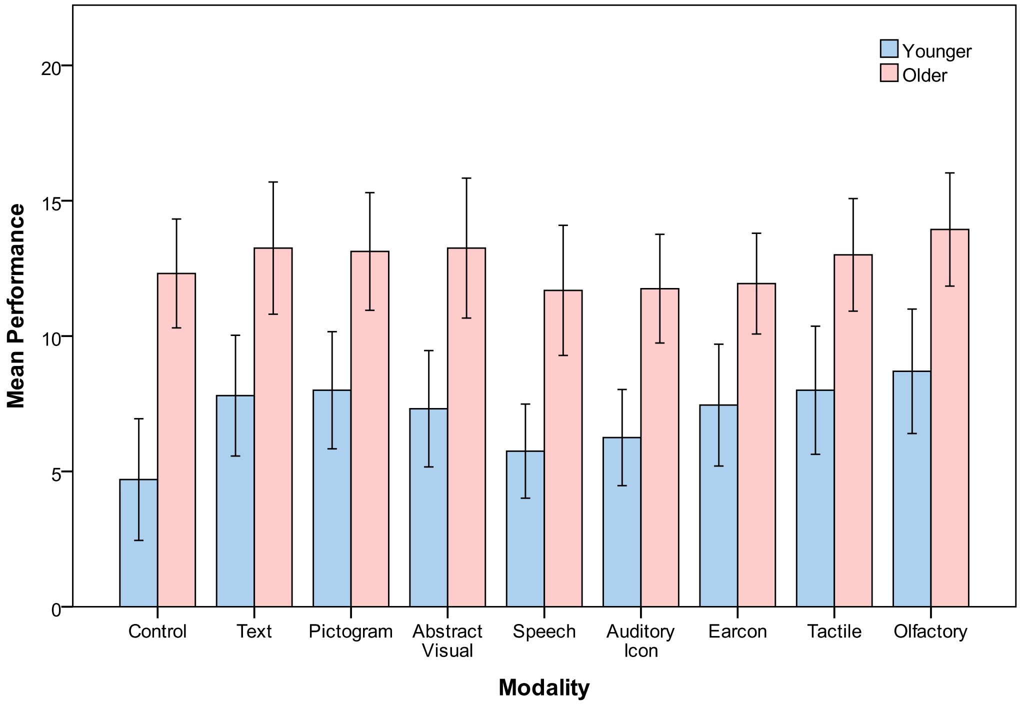
(a) The relationship between modality and temporal demand. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the temporal demand scores for the modalities.

Group		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.000^{***}							
	Pic	0.065	0.898						
	Abs	0.545	0.276	0.985					
	Spe	0.670	0.189	0.958	1.000				
	Aud	0.846	0.091	0.861	1.000	1.000			
	Ear	0.898	0.065	0.800	1.000	1.000	1.000		
	Tac	0.189	0.670	1.000	1.000	0.998	0.978	0.958	
	Olf	0.202	0.650	1.000	1.000	0.998	0.982	0.964	1.000
Older	Tex	0.849							
	Pic	1.000	0.970						
	Abs	0.848	1.000	0.970					
	Spe	1.000	0.987	1.000	0.987				
	Aud	0.949	1.000	0.995	1.000	0.999			
	Ear	1.000	0.980	1.000	0.980	1.000	0.998		
	Tac	0.992	1.000	1.000	1.000	1.000	1.000	1.000	
	Olf	0.003^{**}	0.285	0.013^{*}	0.285	0.020^{*}	0.155	0.016^{*}	0.068

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.5: Additional data for TLX performance.



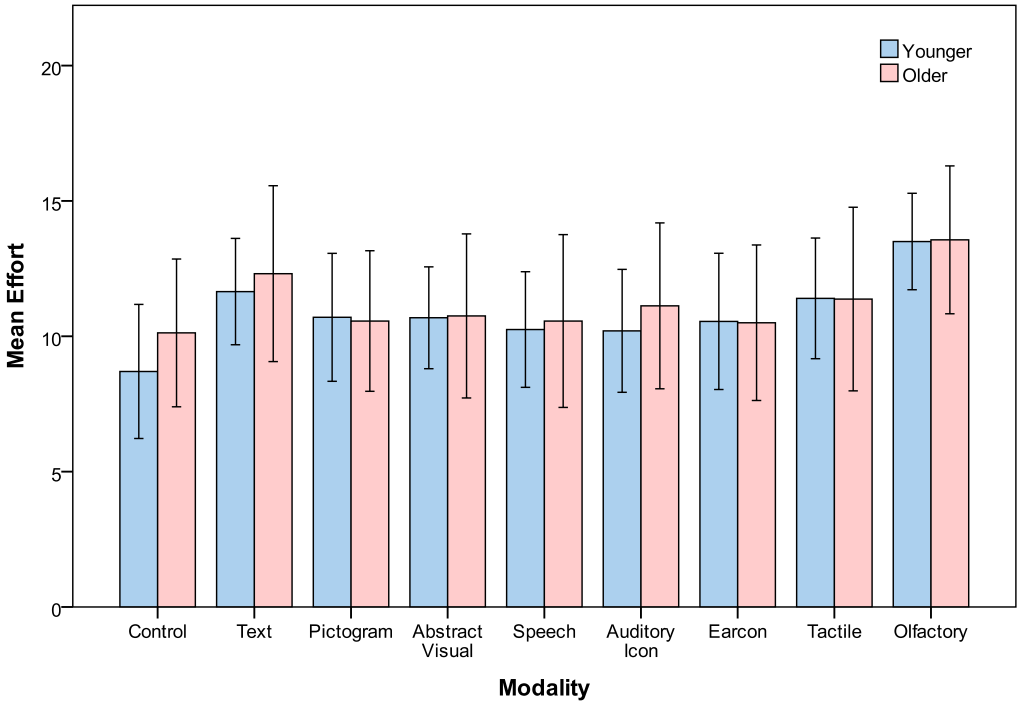
(a) The relationship between modality and performance. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the performance scores for the modalities.

Group		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.015*							
	Pic	0.010*	1.000						
	Abs	0.219	0.991	0.980					
	Spe	0.934	0.423	0.348	0.949				
	Aud	0.797	0.649	0.566	0.992	1.000			
	Ear	0.053	1.000	1.000	1.000	0.689	0.870		
	Tac	0.017*	1.000	1.000	0.992	0.443	0.668	1.000	
	Olf	0.002**	1.000	1.000	0.870	0.147	0.296	0.992	1.000
Older	Tex	0.935							
	Pic	0.992	1.000						
	Abs	0.903	1.000	1.000					
	Spe	0.996	0.463	0.715	0.397				
	Aud	1.000	0.813	0.952	0.757	1.000			
	Ear	1.000	0.952	0.995	0.925	0.993	1.000		
	Tac	0.915	1.000	1.000	1.000	0.419	0.776	0.935	
	Olf	0.315	0.980	0.890	0.989	0.040*	0.169	0.355	0.987

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.6: Additional data for TLX effort.



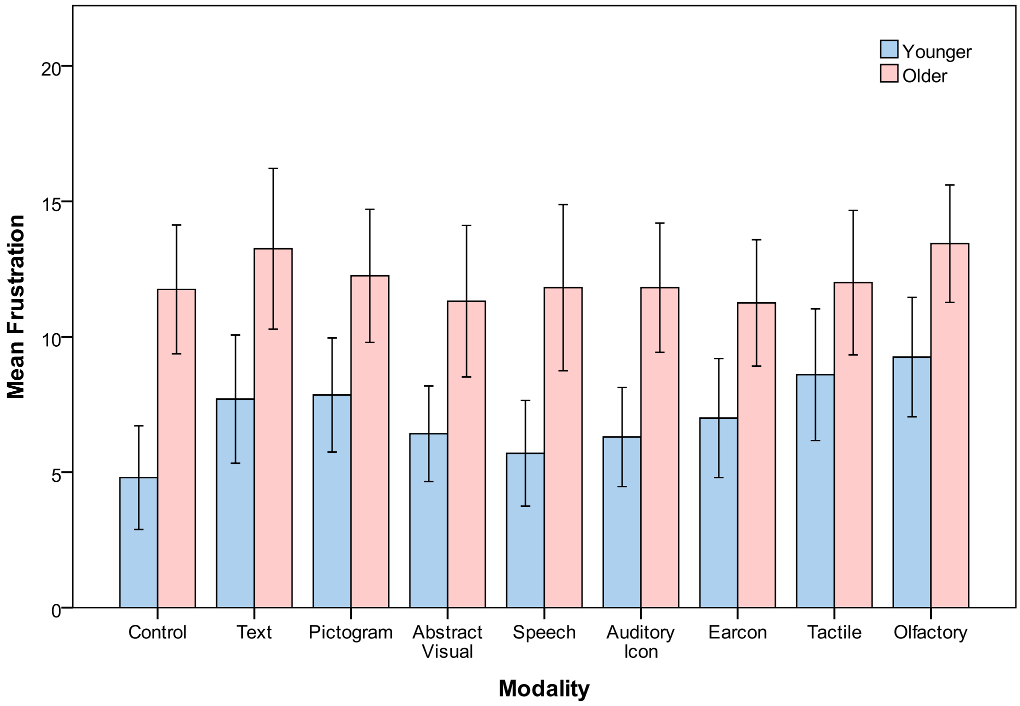
(a) The relationship between modality and effort. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the effort scores for the modalities.

Group		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.016*							
	Pic	0.211	0.993						
	Abs	0.410	0.943	1.000					
	Spe	0.614	0.832	0.999	1.000				
	Aud	0.860	0.573	0.981	0.999	1.000			
	Ear	0.532	0.886	1.000	1.000	1.000	1.000		
	Tac	0.199	0.994	1.000	1.000	0.999	0.977	1.000	
	Olf	0.000***	0.750	0.186	0.077	0.032*	0.008**	0.046*	0.199
Older	Tex	0.333							
	Pic	0.991	0.905						
	Abs	1.000	0.684	1.000					
	Spe	1.000	0.444	0.998	1.000				
	Aud	0.986	0.928	1.000	1.000	0.996			
	Ear	1.000	0.468	0.998	1.000	1.000	0.997		
	Tac	0.589	1.000	0.986	0.892	0.707	0.991	0.729	
	Olf	0.040*	0.995	0.375	0.162	0.066	0.421	0.073	0.947

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure B.7: Additional data for TLX frustration.



(a) The relationship between modality and frustration. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the frustration scores for the modalities.

		Con	Tex	Pic	Abs	Spe	Aud	Ear	Tac
Younger	Tex	0.068							
	Pic	0.021*	1.000						
	Abs	0.490	0.992	0.930					
	Spe	0.997	0.409	0.195	0.939				
	Aud	0.921	0.771	0.511	0.998	1.000			
	Ear	0.636	0.970	0.850	1.000	0.979	1.000		
	Tac	0.006**	0.999	1.000	0.788	0.087	0.300	0.656	
	Olf	0.003**	0.995	1.000	0.696	0.058	0.221	0.552	1.000
Older	NOT SIGNIFICANT								

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Appendix C

Additional Materials – Chapter 6

This appendix presents a full description of the materials used in the third study of this thesis, which is described in Chapter 6.

C.1 Experiment Materials

The NASA-TLX form used was identical to the one used in the first study, which is provided in Appendix A (p. 271). The consent form used was also identical (p. 270). The remaining materials are shown here. The receipts used in the shopping task are provided in the accompanying materials, details of which are provided in Appendix E.

C.2 Raw Ranking Data

This section presents the raw ranking data used in the study. The tables provided here have graphical counterparts in Chapter 6.

C.3 Survey Questions 6-9 Full Analysis

This section presents the full analysis of the ranking data from survey questions 6-9 (difficulty, popularity, suitability for the home and suitability for care) which is presented in Section C.2. These results are summarised relative to their hypotheses in Section 6.6; this appendix presents the full analysis including all factors. The tables provided here have graphical counterparts in Chapter 6.

C.4 NASA-TLX Full Results

Section 4.6 presented the results of the NASA-TLX. The full results of the NASA-TLX were omitted for brevity. Full results are presented here. See also Figure 6.13 on page 237, which may clarify how the overall workload is calculated from the component parts.

Multimodal Reminders in the Home

You are invited to take part in a research study to explore different ways of supporting task organization and memory in the home.

This experiment will take around 90 minutes, and you will be paid £10 to participate. **To participate in this experiment, you must wear jeans, trousers or any other type of legwear with pockets on the day of the experiment.**

You will be asked to carry out 5 activities (sorting clean socks into pairs, taking notes from an audiobook, watching TV, ordering groceries and counting household expenses) over a 15 minute period. In addition, you will need to press 3 buttons at certain points over the 15 minute period to represent cooking, eating and taking medicine.

Your objective is to organize your time effectively between all tasks, and carry out required actions on time. You will be given specific written instructions about when to carry out these tasks before each condition, which you can keep for reference. A clock will be available to you at all times to help you organise your tasks and activities. It will not be possible to complete all the activities in the given timeframe; the objective is instead to balance your time well.

There will be 4 trials during the experiment which you will receive in a random order. In one of the trials, you will have to organise your activities with only the clock for guidance. In the other 3 trials, you will also receive visual, auditory and tactile notifications to assist you. The aim of the experiment is to find out which methods are the most effective at helping participants to organising their time.

At the end of each trial there will be a short survey and at the end of the experiment there will be a final survey and a brief interview. The interview will be recorded (audio only) and transcribed. Only anonymised extracts from the transcripts will be used; the original interview audio files will never be published or made public at any time.

We are interested in finding out about the different types of reminders and how people play the game. You do not need to have technical expertise to take part. In this experiment we are looking for participants between 18-30 years old and participants over 60 years old. If you have a significant visual or audio impairment, you must use corrective technology such as glasses/contact lenses or a hearing aid to take part.

Your participation in this study will be confidential. All tapes and notes will be put in locked filing cabinets: computer data will be stored securely under passwords in compliance with data protection protocols. You will never be identified personally in any reports or publications.

We hope your participation will help us improve the design of computer technology for the home. We aim to improve the effectiveness of reminder technology and to make it more appropriate for use at home. We believe this can be done by using different interaction methods depending on the user's preferences, the importance of the message and the home environment. We hope that you will enjoy learning about new research ideas and technologies.

The research will mainly be carried out by the following people:

- Mr David Warnock, a PhD student at the University of Glasgow
- Dr. Marilyn McGee-Lennon, a researcher at the University of Glasgow

Your participation is voluntary. You can ask us questions about the study before you decide to participate, and you can withdraw from the study at any time. If you withdraw from the study any data collected will be returned to you or destroyed.

To arrange to take part in the study, or if you have any questions, please contact:

David Warnock
Computing Science,
University of Glasgow,
8–17 Lilybank Gardens,
Glasgow, G12 8QQ
Email: warnockd@dcs.gla.ac.uk

You can also visit the MultiMemoHome website for more information on what we do.

http://MultiMemoHome.org

Demographic Questionnaire

Please answer all of the following questions. Circle the answer or write in your answers in the given spaces.

Question 1

What is your gender?

1) Female
2) Male
3) Prefer not to Say

Question 2

What is your age?

____

Question 3

Do you have any sensory (visual, auditory, tactile, olfactory or taste) impairments?

1) No
2) Yes, but corrective technology (glasses, contacts, hearing aid, etc.) not necessary
3) Yes, but it is effectively managed by corrective technology
4) Yes, but it is a problem even with corrective technology *
5) Not sure
6) Rather not say *

If you answered 2 or 5 to the above question, please give details.

____

____

____

* If you have answered '4' or '6' to Question 3, then you may not be able to participate in this experiment.

Post-Experiment Questionnaire

Please answer all of the following questions. Circle the answer or write in your answers in the given spaces.

Question 1

How well did the sock-sorting activity represent something you would do in your own home?

1) Regular household activity
2) Somewhat regular household activity
3) Irregular household activity
4) Rare household activity
5) Never carried out at home.

Can you think of a similar task you regularly carry out at home?

___

Question 2

How well did the radio-listening activity represent something you would do in your own home?

1) Regular household activity
2) Somewhat regular household activity
3) Irregular household activity
4) Rare household activity
5) Never carried out at home.

Can you think of a similar task you regularly carry out at home?

___

Question 3

How well did the TV-watching activity represent something you would do in your own home?

1) Regular household activity
2) Somewhat regular household activity
3) Irregular household activity
4) Rare household activity
5) Never carried out at home.

Can you think of a similar task you regularly carry out at home?

___

Question 4

How well did the budgeting activity represent something you would do in your own home?

1) Regular household activity
2) Somewhat regular household activity
3) Irregular household activity
4) Rare household activity
5) Never carried out at home.

Can you think of a similar task you regularly carry out at home?

___

Question 5

How well did the shopping activity represent something you would do in your own home?

1) Regular household activity
2) Somewhat regular household activity
3) Irregular household activity
4) Rare household activity
5) Never carried out at home.

Can you think of a similar task you regularly carry out at home?

____

Question 6

You took part in four conditions during this experiment. Which condition did you find the easiest? Please rank the conditions, with ‘1’ being the easiest and ‘4’ being the hardest.

1: _____ (easiest)

2: _____

3: _____

4: _____ (hardest)

Question 7

Which condition did you like the best? Please rank the conditions, with ‘1’ being the one you liked most and ‘4’ being the one you liked least.

1: _____ (most liked)

2: _____

3: _____

4: _____ (least liked)

Question 8

Which of the conditions would be the most appropriate for use at home? Please rank the conditions, with ‘1’ being most appropriate for the home and ‘4’ being least appropriate for the home.

1: _____ (most appropriate for home)

2: _____

3: _____

4: _____ (least appropriate for home)

Question 9

Which of the conditions would be the most appropriate for use in a care setting? Please rank the conditions, with ‘1’ being the most appropriate for care and ‘4’ being the least appropriate for care.

1: _____ (most appropriate for care)

2: _____

3: _____

4: _____ (least appropriate for care)

Instructions (1)

You are to imagine that you are at home in the evening with some activities to carry out before going to bed. You know that you cannot complete all the activities before bed, but they are all equally important to you. You have decided to spend your time equally on each task even though it means none of the tasks will be completed by the end of your night.

You have 15 minutes to divide as equally as possible between the following activities:

- Sorting clean socks into matching pairs.
- Listening to the radio.
- Watching some television.
- Writing a shopping list.
- Calculating your household expenditure.

Ideally you want to spend 3 minutes on each of the five activities. Spending less time on an activity is equal to spending more time; try to aim as close to 3 minutes as you can. You can do the activities in any order you wish and can switch activities at will; *however you are not allowed to carry out two activities simultaneously*. To help you organise your time you can view a clock by pressing the button labelled 'clock'. The button is only to let the experimenter known when you are looking at the clock, and you are free to look at the clock as often as you want.

While you try and carry out these activities, you also have to carry out 3 tasks at set times. These tasks are cooking, eating and taking medicine. These tasks should be carried out as follows:

- After 4:30 minutes, you need to start cooking. Do this by pressing the *'cook'* button.
- The food takes time to cook. Press the *'eat'* button 7:00 minutes after you have pressed the *'cook'* button.
- You need to take your medicine on a full stomach. Press the *'medicine'* button 3:15 minutes after pressing the *'eat'* button.

You can keep this instruction card as you carry out your tasks. After 15 minutes the condition will end.

Instructions (2)

You are to imagine that you are at home in the evening with some activities to carry out before going to bed. You know that you cannot complete all the activities before bed, but they are all equally important to you. You have decided to spend your time equally on each task even though it means none of the tasks will be completed by the end of your night.

You have 15 minutes to divide as equally as possible between the following activities:

- Sorting clean socks into matching pairs.
- Listening to the radio.
- Watching some television.
- Writing a shopping list.
- Calculating your household expenditure.

Ideally you want to spend 3 minutes on each of the five activities. Spending less time on an activity is equal to spending more time; try to aim as close to 3 minutes as you can. You can do the activities in any order you wish and can switch activities at will; *however you are not allowed to carry out two activities simultaneously*. To help you organise your time you can view a clock by pressing the button labelled 'clock'. The button is only to let the experimenter known when you are looking at the clock, and you are free to look at the clock as often as you want.

While you try and carry out these activities, you also have to carry out 3 tasks at set times. These tasks are cooking, eating and taking medicine. These tasks should be carried out as follows:

- After 6:45 minutes, you need to start cooking. Do this by pressing the *'cook'* button.
- The food takes time to cook. Press the *'eat'* button 3:00 minutes after you have pressed the *'cook'* button.
- You need to take your medicine on a full stomach. Press the *'medicine'* button 2:30 minutes after pressing the *'eat'* button.

You can keep this instruction card as you carry out your tasks. After 15 minutes the condition will end.

Instructions (3)

You are to imagine that you are at home in the evening with some activities to carry out before going to bed. You know that you cannot complete all the activities before bed, but they are all equally important to you. You have decided to spend your time equally on each task even though it means none of the tasks will be completed by the end of your night.

You have 15 minutes to divide as equally as possible between the following activities:

- Sorting clean socks into matching pairs.
- Listening to the radio.
- Watching some television.
- Writing a shopping list.
- Calculating your household expenditure.

Ideally you want to spend 3 minutes on each of the five activities. Spending less time on an activity is equal to spending more time; try to aim as close to 3 minutes as you can. You can do the activities in any order you wish and can switch activities at will; *however you are not allowed to carry out two activities simultaneously*. To help you organise your time you can view a clock by pressing the button labelled 'clock'. The button is only to let the experimenter known when you are looking at the clock, and you are free to look at the clock as often as you want.

While you try and carry out these activities, you also have to carry out 3 tasks at set times. These tasks are cooking, eating and taking medicine. These tasks should be carried out as follows:

- After 3:30 minutes, you need to start cooking. Do this by pressing the *'cook'* button.
- The food takes time to cook. Press the *'eat'* button 5:00 minutes after you have pressed the *'cook'* button.
- You need to take your medicine on a full stomach. Press the *'medicine'* button 4:45 minutes after pressing the *'eat'* button.

You can keep this instruction card as you carry out your tasks. After 15 minutes the condition will end.

Instructions (4)

You are to imagine that you are at home in the evening with some activities to carry out before going to bed. You know that you cannot complete all the activities before bed, but they are all equally important to you. You have decided to spend your time equally on each task even though it means none of the tasks will be completed by the end of your night.

You have 15 minutes to divide as equally as possible between the following activities:

- Sorting clean socks into matching pairs.
- Listening to the radio.
- Watching some television.
- Writing a shopping list.
- Calculating your household expenditure.

Ideally you want to spend 3 minutes on each of the five activities. Spending less time on an activity is equal to spending more time; try to aim as close to 3 minutes as you can. You can do the activities in any order you wish and can switch activities at will; *however you are not allowed to carry out two activities simultaneously*. To help you organise your time you can view a clock by pressing the button labelled 'clock'. The button is only to let the experimenter known when you are looking at the clock, and you are free to look at the clock as often as you want.

While you try and carry out these activities, you also have to carry out 3 tasks at set times. These tasks are cooking, eating and taking medicine. These tasks should be carried out as follows:

- After 5:15 minutes, you need to start cooking. Do this by pressing the *'cook'* button.
- The food takes time to cook. Press the *'eat'* button 2:45 minutes after you have pressed the *'cook'* button.
- You need to take your medicine on a full stomach. Press the *'medicine'* button 4:30 minutes after pressing the *'eat'* button.

You can keep this instruction card as you carry out your tasks. After 15 minutes the condition will end.

Household Budget

Your task is to work out how much you've spent over the previous month on groceries, clothes, bills, transportation and everything else. Using the receipts and invoices you have been given, add up the total money spent in each of those categories. Use the following workspace and write the final totals at the bottom.

Workspace

CATEGORY	Groceries	Clothing	Household Bills	Transportation	Everything Else
EXPENSE					
TOTAL					

Shopping List

You have a number of events in the upcoming month, and it's desirable for you to get all your shopping done at once. Using the given catalogue, you pick out items to buy that satisfy the following requirements. Write down a suggestion for each scenario below along with the catalogue number and price.

Item 1

You have to get a gift for a friend's flat-warming. This is their first flat so they don't have much furniture or ornaments. You have a budget of £45 to pick out a gift for them.

Item 2

It's your friend's birthday and you want to get them something nice. He likes to cook so something for the kitchen might be nice. He also likes football. You have a budget of £30.

Item 3

A friend's child is having their 7th birthday soon. You want to get them an educational gift with a budget of £25.

Item 4

As a New Year's Resolution you wanted to try and get into better shape. With a budget of £60, find some exercise equipment you can use at home.

Shopping List

You have a number of events in the upcoming month, and it's desirable for you to get all your shopping done at once. Using the given catalogue, you pick out items to buy that satisfy the following requirements. Write down a suggestion for each scenario below along with the catalogue number and price.

Item 1

You have to get a gift for a friend's house-warming. Their new house has an overgrown garden and your friend has been talking excitedly about their plans for it. You have a budget of £35 to pick out a gift for them.

Item 2

It's your friend's birthday and you want to get them something nice. She loves to look nice, so jewelry or beauty products would make a nice gift. You have a budget of £50.

Item 3

A friend's child is having their 9th birthday soon. Your other friends are buying education gifts, so you want to get them something cool and fun. You have a budget of £30.

Item 4

Your friends keep asking you to go jogging with them, but you don't have any exercise clothes. You've set aside £45 to buy something some.

Shopping List

You have a number of events in the upcoming month, and it's desirable for you to get all your shopping done at once. Using the given catalogue, you pick out items to buy that satisfy the following requirements. Write down a suggestion for each scenario below along with the catalogue number and price.

Item 1

A close friend is getting married and you want to get something nice for the happy couple, but they don't have a wedding list and haven't asked for anything. You've set aside £100 for a gift.

Item 2

A friend is leaving to work abroad and you're going to their leaving party soon. You want to get her a gift to remind them of the UK on a budget of £20.

Item 3

Your cousin will be turning 11 soon, and you want to get them an appropriate gift with a budget of £30. However, you don't see your cousin very often so you're not sure what they like.

Item 4

You've decided it's time to buy a new digital camera. Pick out one you like on a budget of £120.

Shopping List

You have a number of events in the upcoming month, and it's desirable for you to get all your shopping done at once. Using the given catalogue, you pick out items to buy that satisfy the following requirements. Write down a suggestion for each scenario below along with the catalogue number and price.

Item 1

A friend has recently passed their driving test and bought their first car. You want to get them an appropriate gift on a budget of £25.

Item 2

Your friend's birthday is coming up. They love golfing and you want to get them something golf-related on a budget of £20.

Item 3

Your a friend's child is having their 5th birthday soon. You want to get them something fun but also educational on a £30 budget.

Item 4

You've been saving up for a while and it's finally time to get a nice TV. You've managed to save up £550. Pick out a flat-panel TV that you like.

Table C.1: Table showing the distribution of ranking data for the familiarity of the home-style tasks. This data was gathered in response to questions 1–5 of the subjective survey shown on page 304. $N = 20$ for the younger group and $N = 19$ for the older group.

Group	Activity	Regular	Somewhat Regular	Irregular	Rare	Never
Younger	Socks	7	6	4	2	1
	Radio	6	3	2	6	3
	TV	11	5	0	3	1
	Budget	2	3	4	3	8
	Shopping	4	2	5	7	2
Older	Socks	6	6	2	2	1
	Radio	13	1	0	3	0
	TV	8	4	1	1	3
	Budget	8	5	2	2	0
	Shopping	2	6	1	6	2

Table C.2: Table showing the distribution of ranking data for the ease-of-use subjective feedback. This data was gathered in response to question 6 of the subjective survey shown on page 304. $N = 20$ for the younger group and $N = 19$ for the older group.

Group	Condition	1	2	3	4
Younger	No Reminders	1	1	4	14
	Static	12	2	2	4
	Random	1	6	13	0
	Dynamic	6	11	1	2
Older	No Reminders	4	5	0	10
	Static	11	6	2	0
	Random	0	5	13	1
	Dynamic	3	6	3	7

Table C.3: Table showing the distribution of ranking data for the popularity subjective feedback. This data was gathered in response to question 7 of the subjective survey shown on page 304. $N = 20$ for the younger group and $N = 19$ for the older group.

Group	Condition	1	2	3	4
Younger	No Reminders	3	2	5	10
	Static	6	4	5	5
	Random	4	6	7	3
	Dynamic	7	8	3	2
Older	No Reminders	2	5	3	9
	Static	7	6	4	2
	Random	3	6	6	4
	Dynamic	6	5	5	3

Table C.4: Table showing the distribution of ranking data for the 'appropriateness for use at home' subjective feedback. This data was gathered in response to question 8 of the subjective survey shown on page 304. $N = 20$ for the younger group and $N = 19$ for the older group.

Group	Condition	1	2	3	4
Younger	No Reminders	1	5	6	8
	Static	8	3	3	6
	Random	4	7	5	4
	Dynamic	6	6	6	2
Older	No Reminders	5	0	1	13
	Static	9	5	2	3
	Random	0	6	12	1
	Dynamic	5	8	4	2

Table C.5: Table showing the distribution of ranking data for the 'appropriateness for use in a care setting' subjective feedback. This data was gathered in response to question 9 of the subjective survey shown on page 304. $N = 20$ for the younger group and $N = 19$ for the older group.

Group	Condition	1	2	3	4
Younger	No Reminders	0	0	6	14
	Static	9	4	5	2
	Random	4	4	8	4
	Dynamic	7	12	1	0
Older	No Reminders	1	0	0	18
	Static	7	9	3	0
	Random	1	2	16	0
	Dynamic	10	8	0	1

Table C.6: Tables showing the full analysis of the ranked difficulty data. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

(a) Within-groups analysis showing 2-way Freidman's ANOVA for each group.

Group	X^2	Kendall's W	N	df	p
Younger	21.30	0.36	20	3	$<.001^{***}$
Older	14.47	0.25	19	3	$.002^{**}$

(b) Pairwise comparison of within-groups results. Table shows T values annotated to show significance (p) values with Bonferroni correction.

Group		No Reminders	Static	Random	Dynamic
Younger	No Reminders		-1.65^{***}	-0.95	-1.60^{**}
	Static	1.65^{***}		-0.70	-0.05
	Random	0.95	0.70		-0.65
	Dynamic	1.60^{**}	0.05	0.65	
Older	No Reminders		-1.32^{*}	-0.05	-0.11
	Static	1.32^{*}		-1.26^{*}	-1.21^{*}
	Random	0.05	1.26^{*}		-0.50
	Dynamic	0.11	1.21^{*}	0.50	

(c) Between-groups tests for each of the conditions. p_{adj} shows p with the Bonferroni correction applied.

Condition	U	Wilcoxon W	N	Z	p	p_{adj}	r
No Reminders	139.50	329.50	39	-1.63	0.10	0.42	-0.26
Static	174.00	364.00	39	-0.51	0.61	1.00	-0.08
Random	164.50	374.50	39	-0.87	0.39	1.00	-0.14
Dynamic	116.50	326.50	39	-2.19	0.03	0.12	-0.35

Table C.7: Tables showing the full analysis of the ranked preference data. As the analysis was not significant, no pairwise comparisons are provided. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

(a) Within-groups analysis showing 2-way Freidman's ANOVA for each group.

Group	X^2	Kendall's W	N	df	p
Younger	7.38	0.12	20	3	0.06
Older	6.13	0.11	19	3	0.11

(b) Between-groups tests for each of the conditions. p_{adj} shows p with the Bonferroni correction applied.

Condition	U	Wilcoxon W	N	Z	p	p_{adj}	r
No Reminders	180.50	370.50	39	−0.29	0.78	1.00	−0.05
Static	154.00	344.00	39	−1.05	0.29	1.00	−0.17
Random	177.00	387.00	39	−0.38	0.70	1.00	−0.06
Dynamic	164.50	374.50	39	−0.75	0.45	1.00	−0.12

Table C.8: Tables showing the full analysis of the ranked 'suitability for the home' data. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

(a) Within-groups analysis showing 2-way Freidman's ANOVA for each group.

Group	X^2	Kendall's W	N	df	p
Younger	5.26	0.09	20	3	0.154
Older	10.39	0.18	19	3	0.016*

(b) Pairwise comparison of within-groups results for the older participants, as the younger participants were not significant in Table C.8a. Table shows T values annotated to show significance (p) values with Bonferroni correction.

Group		No Reminders	Static	Random	Dynamic
Older	No Reminders		−1.21*	−0.42	−1.00
	Static	1.21*		−0.79	−0.21
	Random	0.42	0.79		−0.58
	Dynamic	1.00	0.21	0.58	

(c) Between-groups tests for each of the conditions. p_{adj} shows p with the Bonferroni correction applied.

Condition	U	Wilcoxon W	N	Z	p	p_{adj}	r
No Reminders	160.50	370.50	39	−0.91	0.36	1.00	−0.15
Static	159.50	349.50	39	−0.91	0.37	1.00	−0.15
Random	155.00	365.00	39	−1.05	0.29	1.00	−0.17
Dynamic	185.00	375.00	39	−0.15	0.88	1.00	−0.02

Table C.9: Tables showing the full analysis of the ranked 'suitability for care' data. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

(a) Within-groups analysis showing 2-way Freidman's ANOVA for each group.

Group	X^2	Kendall's W	N	df	p
Younger	28.08	0.47	20	3	$<.001^{***}$
Older	36.92	0.65	19	3	$<.001^{***}$

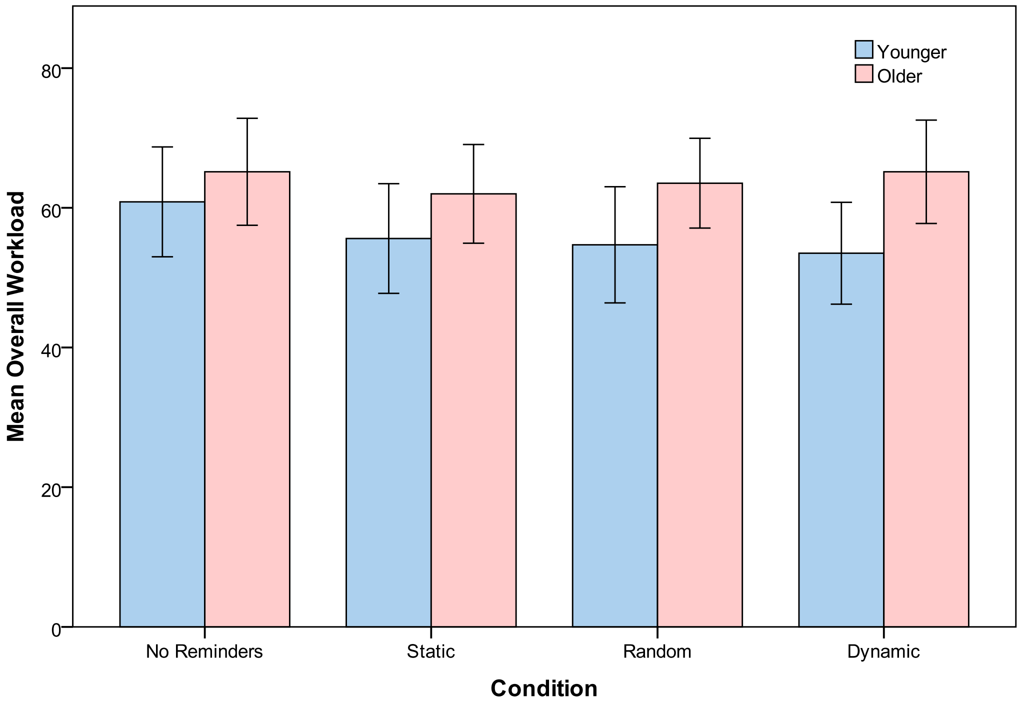
(b) Pairwise comparison of within-groups results. Table shows T values annotated to show significance (p) values with Bonferroni correction.

Group		No Reminders	Static	Random	Dynamic
Younger	No Reminders		-1.70^{***}	-1.10^{*}	-2.00^{***}
	Static	1.70^{***}		-0.60	-0.30
	Random	1.10^{*}	0.60		-0.90
	Dynamic	2.00^{***}	0.30	0.90	
Older	No Reminders		-2.05^{***}	-1.05	-2.26^{***}
	Static	2.05^{***}		-1.00	-0.21
	Random	1.05	1.00		-1.21^{*}
	Dynamic	2.26^{***}	0.21	1.21^{*}	

(c) Between-groups tests for each of the conditions. p_{adj} shows p with the Bonferroni correction applied.

Condition	U	Wilcoxon W	N	Z	p	p_{adj}	r
No Reminders	146.00	356.00	39	-1.86	0.06	0.26	-0.30
Static	177.00	367.00	39	-0.39	0.70	1.00	-0.06
Random	174.00	384.00	39	-0.52	0.61	1.00	-0.08
Dynamic	159.00	349.00	39	-9.84	0.33	1.00	-1.58

Figure C.1: Additional data for TLX overall workload.



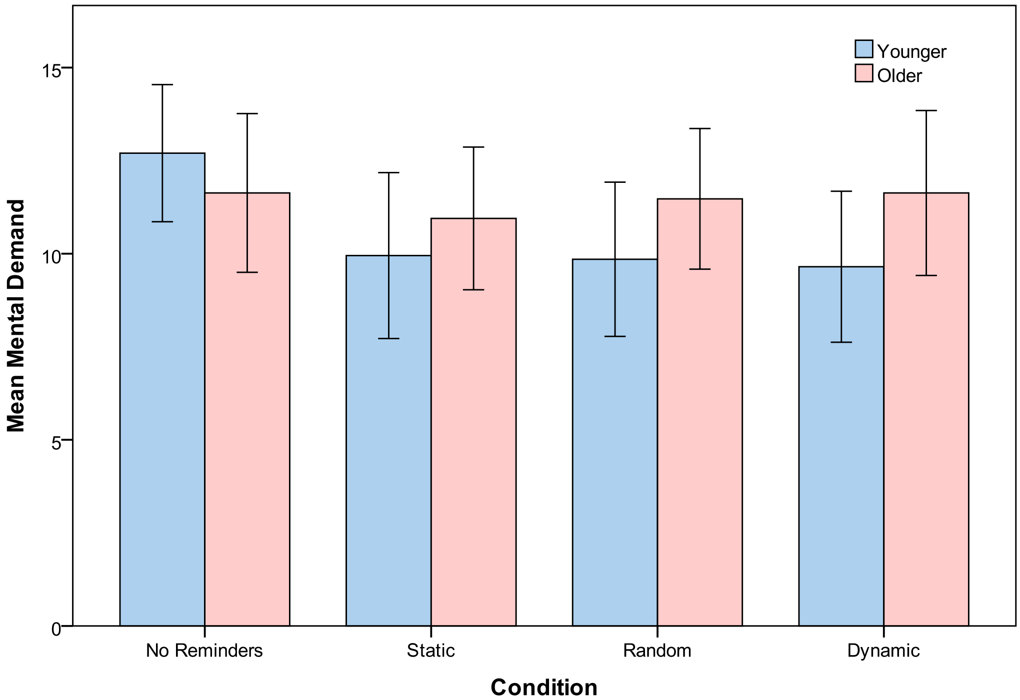
(a) The relationship between modality and overall workload. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the overall workload scores for the modalities.

Group		None	Static	Random	Dynamic
Younger	None		0.046*	0.033*	0.002**
	Static	0.046*		0.999	0.755
	Random	0.033*	0.999		0.822
	Dynamic	0.002**	0.755	0.822	
Older	NOT SIGNIFICANT				

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.2: Additional data for TLX mental demand.



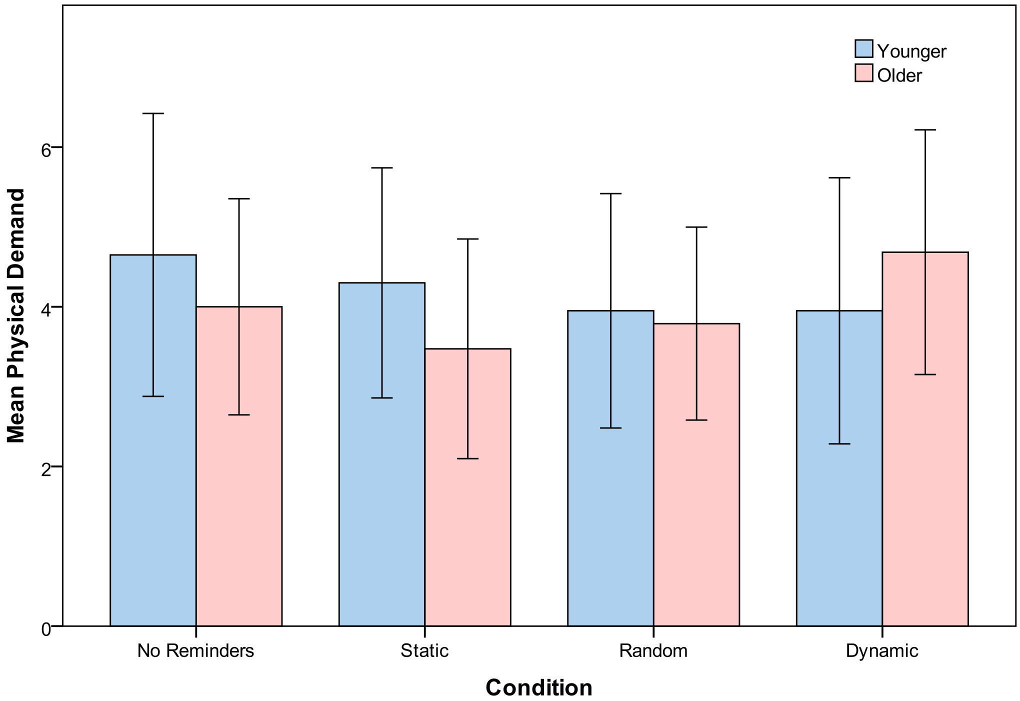
(a) The relationship between modality and mental demand. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the mental demand scores for the modalities.

Group		None	Static	Random	Dynamic
Younger	None		0.003^{**}	0.001^{***}	0.000^{***}
	Static	0.003^{**}		0.982	0.849
	Random	0.001^{***}	0.982		0.972
	Dynamic	0.000^{***}	0.849	0.972	
Older	NOT SIGNIFICANT				

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.3: Additional data for TLX physical demand.



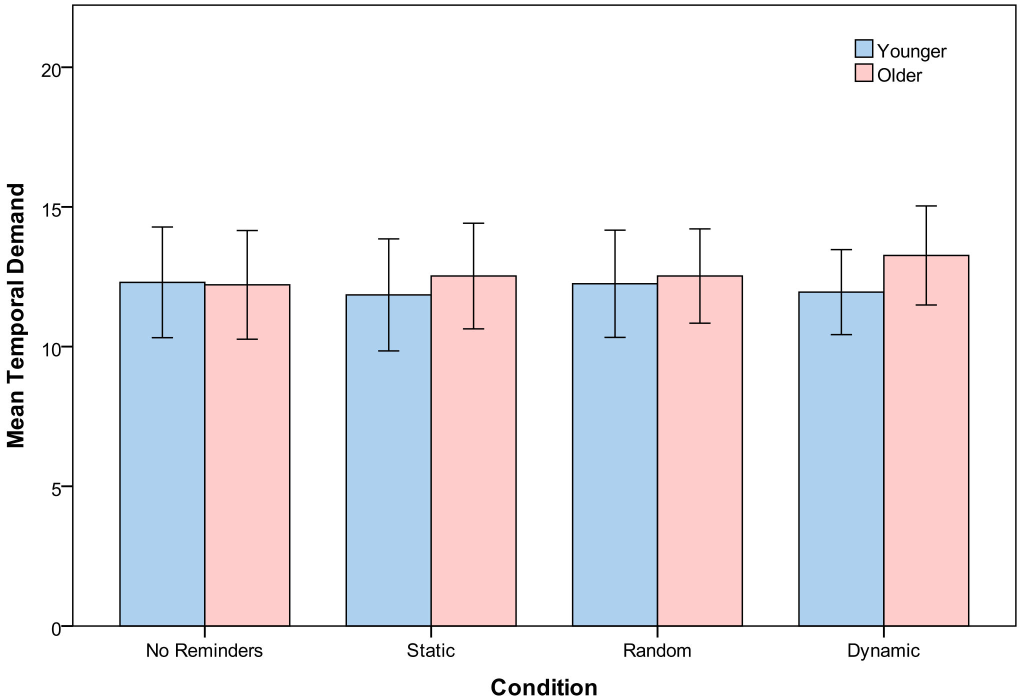
(a) The relationship between modality and physical demand. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the physical demand scores for the modalities.

Group		None	Static	Random	Dynamic
Younger		NOT SIGNIFICANT			
Older	None		0.259	0.709	0.871
	Static	0.259		0.871	0.047*
	Random	0.709	0.871		0.259
	Dynamic	0.871	0.047*	0.259	

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.4: Additional data for TLX temporal demand.



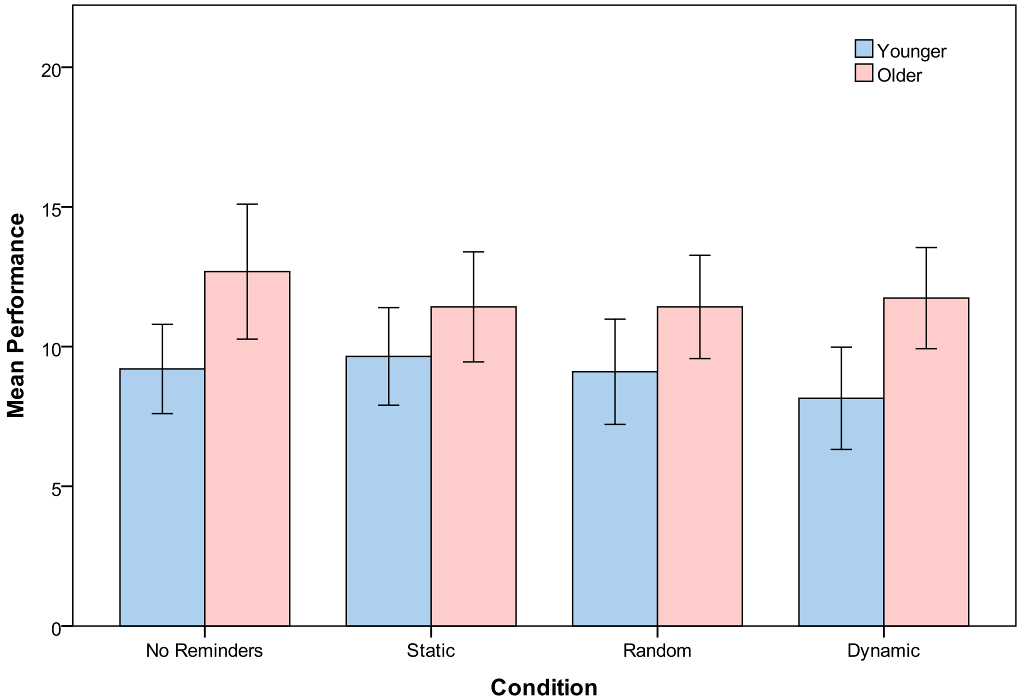
(a) The relationship between modality and temporal demand. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the temporal demand scores for the modalities.

Group	None	Static	Random	Dynamic
Younger	NOT SIGNIFICANT			
Older	NOT SIGNIFICANT			

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.5: Additional data for TLX performance.



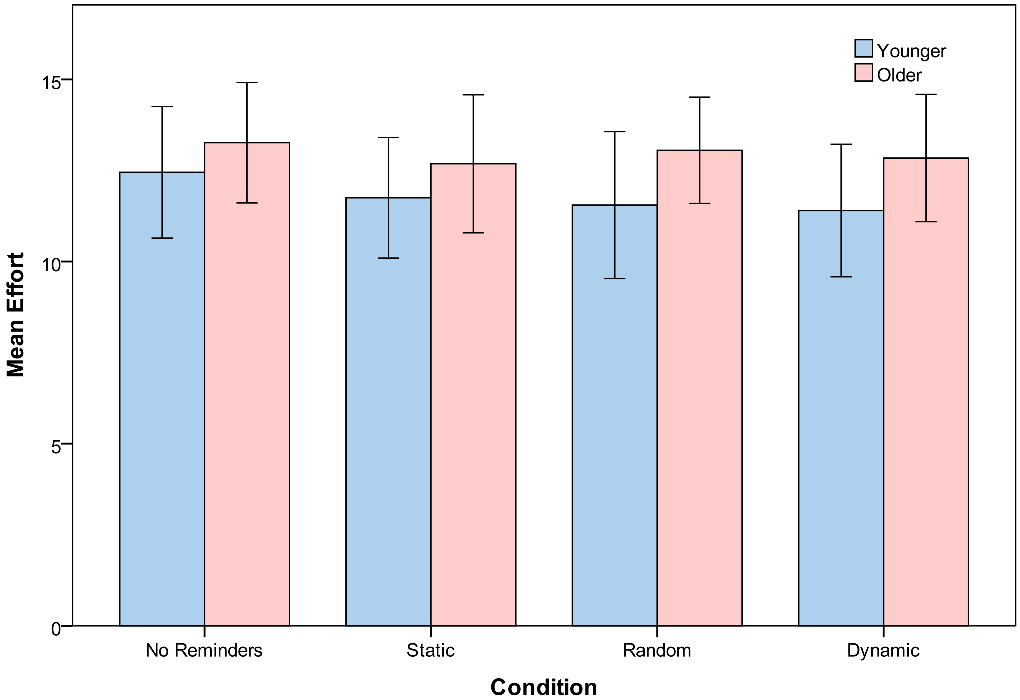
(a) The relationship between modality and performance. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the performance scores for the modalities.

Group	None	Static	Random	Dynamic
Younger	NOT SIGNIFICANT			
Older	NOT SIGNIFICANT			

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.6: Additional data for TLX effort.



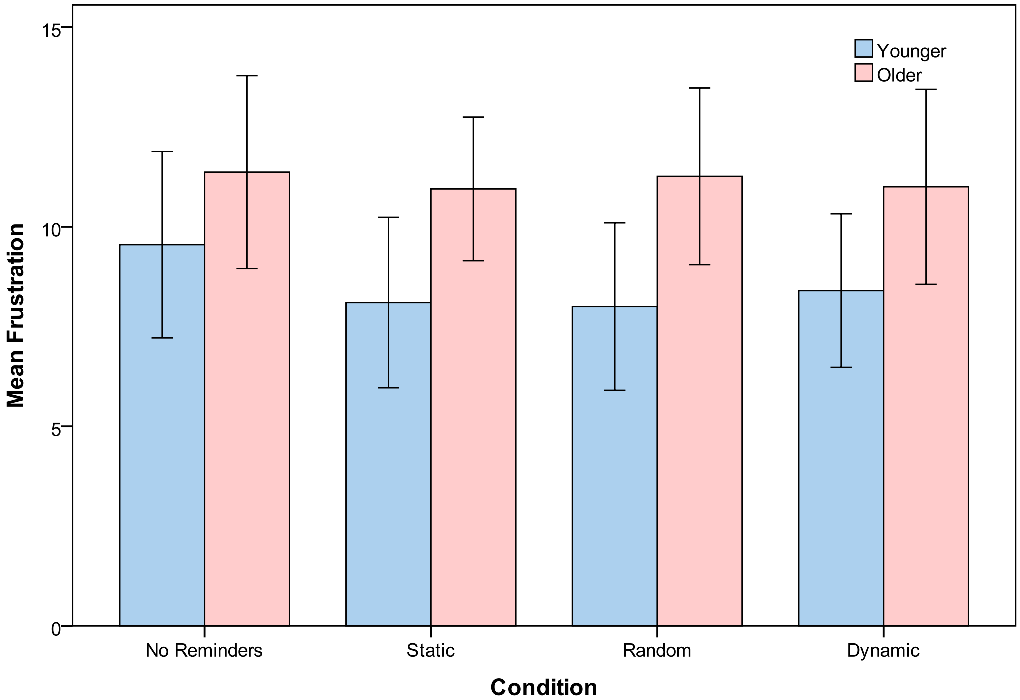
(a) The relationship between modality and effort. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the effort scores for the modalities.

Group	None	Static	Random	Dynamic
Younger	NOT SIGNIFICANT			
Older	NOT SIGNIFICANT			

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Figure C.7: Additional data for TLX frustration.



(a) The relationship between modality and frustration. Error bars show 95% confidence intervals.

(b) Significance (p) values of *post hoc* pairwise comparisons between the frustration scores for the modalities.

Group		None	Static	Random	Dynamic
Younger	None		0.074	0.027*	0.194
	Static	0.074		0.982	0.973
	Random	0.027*	0.982		0.852
	Dynamic	0.194	0.973	0.852	
Older	NOT SIGNIFICANT				

Note: Significance values calculated by R using the method developed by Galili [66], which uses the Tukey method to compensate for the familywise error rate. [*] = Significant to $p < .05$, [**] = Significant to $p < .01$ and [***] = Significant to $p < .001$.

Appendix D

Statistical Notes

This appendix provides additional notes on some of the statistics carried out in the thesis. All statistics in this thesis were performed using IBM SPSS Statistics 20.

D.1 Repeated-Measures General Linear Model (GLM)

Repeated-measures hypotheses were tested using a Repeated-Measures General-Linear Model, which Field [63, 64] also calls GLM4. This test was used repeatedly in Chapters 4 and 6. The primary assumption of this test is that the data has sphericity (denoted ϵ), which means that the that "the level of dependence between experimental conditions is roughly equal" [63, p. 459].

The assumption of sphericity is tested with Mauchley's test. If Mauchley's test is significant, then the assumption of sphericity has been violated and a correction must be applied. Several different types of correction can be applied, but the most common correction in this thesis was the *Greenhouse-Geisser Correction*; Field states that this should be applied when Mauchley's test finds sphericity is less than 0.75 ($\epsilon < .75$). If sphericity is greater than 0.75, then another correction should be applied called the *Huynh-Feldt Correction.*

When carrying out a GLM4 in SPSS, the standard and corrected values are automatically generated as part of the analysis process. Selecting the appropriate result is a simple matter of checking the result of Mauchley's test and using the uncorrected (Mauchley's test is not significant) or corrected (if the assumption of sphericity has been violated) result from the SPSS output.

D.2 Graphing Repeated-Measures Error Bars

SPSS was used to generate all the graphs included in this thesis. However, SPSS does not correctly compute error bars for repeated-measures or mixed-models experimental designs, although it is capable of producing this data as part of the output from the relevant statistical tests.

Instructions for normalising the data that will result in correct error bars are provided in Andy Field's book Discovering Statistics using SPSS [63, p. 316-324]. The method described is the Cousineau method [47], which requires ensuring that all the subjects have the same average. The error bars for all graphs were calculated using variants of the following SPSS Syntax script, with the following code showing the normalisation process specifically for Figure 3.14.

```
AGGREGATE
   /OUTFILE=* MODE=ADDVARIABLES OVERWRITE=YES
   /BREAK=participant_id
   /avg_per_participant = MEAN(avg_matches).
AGGREGATE
   /OUTFILE=* MODE=ADDVARIABLES OVERWRITE=YES
   /grand_mean = MEAN(avg_matches).
EXECUTE.
COMPUTE avg_matches_normal
   =avg_matches - avg_per_participant + grand_mean.
EXECUTE.
VARIABLE LABELS
avg_matches_normal "Matches Per Game".
DELETE VARIABLES avg_per_participant grand_mean.
EXECUTE.
```

For mixed-models designs, such as those found in Chapter 4, each group calculated its own 'grand mean', as shown in the code below (used to generate Figure 4.7). In both cases it was possible to compare the result of this to the output of the statistical processes, which ensured that the resulting error bars were indeed accurate.

```
AGGREGATE
   /OUTFILE=* MODE=ADDVARIABLES OVERWRITE=YES
   /BREAK=experiment_id pid
   /avg_per_participant = MEAN(cards_matched).
AGGREGATE
   /OUTFILE=* MODE=ADDVARIABLES OVERWRITE=YES
   /BREAK=experiment_id
   /grand_mean = MEAN(cards_matched).
EXECUTE.
COMPUTE cards_matched_normal
   = cards_matched - avg_per_participant + grand_mean.
```

```
EXECUTE.
VARIABLE LABELS
cards_matched_normal "Matches Per Game".
DELETE VARIABLES avg_per_participant grand_mean.
EXECUTE.
```

D.3 Non-Parametric Mixed Designs ANOVA

In Chapters 4 and 6, the non-parametric NASA-TLX data would have ideally been processed using a *robust* mixed designs ANOVA. Robust methods use a technique called bootstrapping to resample non-parametric data allowing standard statistical methods to be used. This technique is uncommon so this appendix explains how this technique was attempted and why the results were not included in the main Thesis.

As noted by Andy Field in *Discovering Statistics with SPSS* [63], there is no non-parametric equivalent to the mixed designs ANOVA. However, there are several resampling methods (*i.e.* bootstrapping) that can be used to make the data suitable for use with such methods. While SPSS supports bootstrapping for some methods, it does not support bootstrapping for mixed designs[1].

To carry out non-parametric mixed design ANOVAs the R[2] statistical package must instead be used with the `WRS`[3] package installed. The `WRS` package was created by Rand Wilcox in conjunction with his book *Introduction to Robust Estimation & Hypothesis Testing* [220].

Rand Wilcox can be considered a reliable source on the subject of robust methods. He has a PhD in psychometrics and is a professor of psychology at the University of California. He has published 4 books on statistics of which he is the sole author and over 130 journal articles on statistics and robust methods. He is a Fellow of the Royal Statistics Society and American Psychological Society.[4] His book *Introduction to Robust Estimation & Hypothesis Testing* [220] is cited by Field *et al.* [63, 64] as the most significant work on the subject of robust testing.

The robust mixed design ANOVA was carried out using the methods described by Wilcox [220] and Field *et al.* [64]. The example code provided shows the analysis of the TLX data from Chapter 6.

[1] At the time of writing, current release of SPSS is version 20.
[2] The R Project for Statistical Computing, `http://r-project.org`
[3] Wilcox's Robust Statistics, `http://r-forge.r-project.org/R/?group_id=468`.
[4] Source: University of South California, `http://tinyurl.com/phoxa6e`.

The TLX data contains 6 dependent variables (the tlx measures themselves), 9 repeated-measures variables ($n = 9$ in Chapter 4) and 1 between-groups variable (age group). This was imported into R in long format, with one row per condition.[5]

The data was reshaped into a format where the first n columns are for condition 1 and the second n columns for condition 2, as shown below.

Younger Group			Older Group		
Condition 1	Condition 2	Condition 3	Condition 1	Condition 2	Condition 3

This was done using the `reshape` package. If the groups are not even (true in both analyses) then blank entries will be included in the result, but this will not effect the analysis.

Wilcox [220] outlined several options for carrying out the analysis. Field *et al.* [64] suggests the bootstrap-t method, which is one of the simplest methods of bootstrapping (the primary alternative is called the bootstrap percentage method). The bootstrap-t method involves calculating the trimmed mean (r) by removing outliers from the data then calculating the mean. Running the robust mixed design in R using the bootstrap-t method involves calling the `tsplit` method. One of the parameters for this method is the 'trimming value', or how far an outlier should be before it is trimmed. Both Wilcox [220, p. 411] and Field [64, p. 647] recommend 0.2 for this value.

The result of the test is a test statistic ($\hat{\Psi}$) and a significance test (p value). A result is produced for the between-groups variable, the within-groups variable, and again for interactions between the two. According to Wilcox, this test can only be used with a "between-by-within' design": separate methods are provided for alternative designs.

There are several methods provided by Wilcox for calculating pairwise comparisons depending on the levels of the factors. As the design of the experiments in Chapters 4 and 6 both have a 2-level between-groups variable, the only pairwise comparisons that are needed are on the repeated measures variable. This was done using the bootstrap-t method as before with the trimming value set to 0.2.

This produced two tables, one for each level of the between-groups variable (*i.e.* one younger, one older). This table provides the K test statistic, $\hat{\Psi}$ and significance tests (p values). It also includes a value called p_{crit}, which should be taken as the significance level to beat. This value is generated by the test to help avoid type I errors.

The R code for the entire process was:

```
library(reshape)
```

[5]So the total number of rows is $p * n$, where p is the total number of participants and n was the number of conditions.

```
library(WRS)

headers = c("tlx_mental","tlx_physical",
            "tlx_temporal","tlx_performance",
            "tlx_effort","tlx_frustration")
tlxData<-read.delim("TLX_Data.csv", sep=",", header=TRUE)

for(i in headers){
  print(i)
  tlx_processing<-cast(tlxData, row~group+condition,value=i)
  tlx_processing$row<-NULL
  print(tsplit(2,4,tlx_processing,tr=.2))
  bwbmcp(2,4,tlx_processing,tr=.2,dif=F)
}
```

However, the results of this analysis bore no resemblance to the results included in Chapter 6 of this thesis. Attempting the alternative method described above caused the `WRS` package to throw an internal error, despite the other code running without problems (although it produced surprising output). There is also a minor error in the Wilcox book [220, p. 442] where the section relating to these methods is repeated twice. It is likely that these statistical methods and their corresponding R programs have not yet reached maturity, and as such this method was not used to analyse the non-parametric data presented in this thesis. The code and data to run this analysis is provided in the accompanying materials under `Study 3/Analysis/R_TLX`.

Appendix E

Description of Accompanying Material

This appendix provides a full description of the material that accompanies this thesis. The accompanying material is stored on a CD-ROM and includes source code, raw data and statistical analyses. Each section of this appendix corresponds to a directory on the accompanying CD-ROM and describes the material that can be found inside that directory.

E.1 Study 1

This directory contains three subdirectories as follows:

Analysis This folder contains the SPSS files and `.csv` files used for the statistics in Chapter 3.

Data Contains a complete MySQL dump of the experimental data.

Design Files Contains additional photographs and documents relating to the experiment, including the ethics documentation.

Software Contains an archive containing the Java software used to run the experiment for the first study.

E.2 Study 2

This directory contains three subdirectories as follows:

Analysis This folder contains the SPSS files and `.csv` files used for the statistics in Chapter 4.

Data Contains a complete MySQL dump of the experimental data.

Design Files Contains additional photographs and documents relating to the experiment, including the ethics documentation.

Software Contains an archive containing the Java software used to run the experiment for the second study.

E.3 Study 3

This directory contains three subdirectories as follows:

Analysis This folder contains the SPSS files and `.csv` files used for the statistics in Chapter 6.

Data Contains a complete MySQL dump of the experimental data.

Design Files Contains additional photographs and documents relating to the experiment, including the ethics documentation. This also includes the fake receipts that were omitted from Appendix C.

Software Contains an archive containing the Java software used to run the study, including the Dyna-Cue prototype and the code that was used on the Android device to deliver tactile vibrations.

E.4 Extra Material

This directory contains excel spreadsheets and other loosely related files of relevance to the work presented in this thesis.

E.5 Papers

This directory contains `.pdf` copies of all the papers published based on the work presented in this thesis, along with a `.bbl` BibTEXfile.

Bibliography

[1] Adams, A., and Edworthy, J. Quantifying and predicting the effets of basic text display variables on the perceived urgency of warning labels: tradeoffs involving font size, border weight and colour. *Ergonomics 38*, 11 (Nov. 1995), 2221–2237.

[2] Altmann, E. M., and Trafton, J. G. Memory for goals: an activation-based model. *Cognitive Science 26*, 1 (Jan. 2002), 39–83.

[3] Altmann, E. M., and Trafton, J. G. Task interruption: Resumption lag and the role of cues. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (Westin River North, Chicago, USA, 2004), CogSci 2004, pp. 43–48.

[4] Altmann, E. M., and Trafton, J. G. Timecourse of recovery from task interruption: data and a model. *Psychonomic Bulletin & Review 14*, 6 (Dec. 2007), 1079–1084.

[5] Arrabito, G. R., Mondor, T., and Kent, K. Judging the urgency of non-verbal auditory alarms: a case study. *Ergonomics 47*, 8 (June 2004), 821–40.

[6] Arroyo, E., Bonanni, L., and Selker, T. Waterbot: exploring feedback and persuasive techniques at the sink. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA, 2005), CHI '05, ACM Press, pp. 631–639.

[7] Arroyo, E., and Selker, T. Arbitrating multimodal outputs: Using ambient displays as interruptions. In *Procedings of Human-Computer Interaction: Theory and PRactice* (2003), pp. 591–595.

[8] Arroyo, E., and Selker, T. Self-adaptive multimodal-interruption interfaces. In *Proceedings of the 8th international conference on Intelligent user interfaces* (Miami, Florida, USA, 2003), IUI '03, ACM Press, pp. 6–11.

[9] Arroyo, E., Selker, T., and Stouffs, A. Interruptions as multimodal outputs: which are the less disruptive? In *Proceedings. Fourth IEEE International*

Conference on Multimodal Interfaces (Pittsburgh, Pennsylvania, USA, 2002), ICMI '02, IEEE, pp. 479–482.

[10] Arroyo, E., Sullivan, S., and Selker, T. CarCoach: A polite and effective driving coach. In *CHI '06 extended abstracts on Human factors in computing systems* (Montreal, Quebec, Canada, 2006), CHI EA '06, ACM Press, pp. 357–362.

[11] Bailey, B. P., and Konstan, J. A. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior 22*, 4 (July 2006), 685–708.

[12] Bailey, B. P., Konstan, J. A., and Carlis, J. V. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *INTERACT 2001* (2001), vol. 1, Citeseer, pp. 593–601.

[13] Baker-Ward, L., and Ornstein, P. A. Age differences in visual-spatial memory performance: do children really out-perform adults when playing Concentration? *Bulletin of the Psychonomic Society 26*, 4 (1988), 331–332.

[14] Bardram, J. E., and Hansen, T. R. The AWARE architecture: supporting context-mediated social awareness in mobile cooperation. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (2004), CSCW '04, ACM Press, pp. 192–201.

[15] Barsch, J. Barsch Learning Style Inventory, 1991.

[16] Bavelier, D., and Neville, H. J. Cross-modal plasticity: where and how? *Nature Reviews Neuroscience 3*, 6 (June 2002), 443–52.

[17] Beaudin, J. S., Intille, S. S., and Morris, M. E. To Track or Not to Track: User Reactions to Concepts in Longitudinal Health Monitoring. *Journal of Medical Internet Research 8*, 4 (Dec. 2006), e29.

[18] Beaudin, J. S., Intille, S. S., and Tapia, E. M. Lessons learned using ubiquitous sensors for data collection in real homes. In *Extended abstracts of the 2004 conference on Human factors and computing systems* (Vienna, Austria, 2004), CHI '04, ACM Press, p. 1359.

[19] Bell, G., and Dourish, P. YesterdayâĂŹs tomorrows: notes on ubiquitous computingâĂŹs dominant vision. *Personal and Ubiquitous Computing 11*, 2 (2007), 133–143.

[20] Berg, E. A. A Simple Objective Technique for Measuring Flexibility in Thinking. *Journal of General Psychology 39*, 1 (1948), 15–22.

[21] Berg, W. P., Alessio, H. M., Mills, E. M., and Tong, C. Circumstances

and consequences of falls in independent community-dwelling older adults. *Age and Ageing 26*, 4 (1997), 261–268.

[22] Bickmore, T., Mauer, D., Crespo, F., and Brown, T. Persuasion, Task Interruption and Health Regimen Adherence. In *Second International Conference on Persuasive Technology* (Palo Alto, CA, USA, 2007), vol. 4744 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 1–11.

[23] Bickmore, T., Mauer, D., Crespo, F., and Brown, T. Negotiating Task Interruptions with Virtual Agents for Health Behavior Change (Short Paper). In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems* (2008), vol. 3 of *AAMAS '08*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1241–1244.

[24] Blythe, M., and Dearden, A. Representing older people: towards meaningful images of the user in design scenarios. *Universal Access in the Information Society 8*, 1 (June 2009), 21–32.

[25] Bodnar, A., Corbett, R., and Nekrasovski, D. AROMA: ambient awareness through olfaction in a messaging application. In *Proceedings of the 6th international conference on Multimodal interfaces* (State College, PA, USA, 2004), ICMI '04, ACM, pp. 183–190.

[26] Bogunovich, P., and Salvucci, D. D. The effects of time constraints on user behavior for deferrable interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada, 2011), CHI '11, ACM, pp. 3123–3126.

[27] Bonanni, L., Arroyo, E., Lee, C.-H., and Selker, T. Smart sinks: real-world opportunities for context-aware interaction. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, Oregon, USA, 2005), CHI EA '05, ACM Press, pp. 1232–1235.

[28] Boyce, J., and Shone, G. Effects of Ageing on Smell and Taste. *Postgraduate Medical Journal 82*, 966 (2006), 239–241.

[29] Brehmer, M., McGrenere, J., Tang, C., and Jacova, C. Investigating interruptions in the context of computerised cognitive testing for older adults. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (Austin, Texas, USA, 2012), CHI '12, ACM Press, pp. 2649–2658.

[30] Brewster, S., and Brown, L. M. Tactons: structured tactile messages for non-visual information display. In *Proceedings of the 5th Australasian User Interface Conference* (Dunedin, New Zealand, 2004), A. Cockburn, Ed., vol. 28 of *AUIC '04*, Australian Computer Society, Inc., pp. 15–23.

[31] Brewster, S., McGookin, D., and Miller, C. Olfoto: designing a smell-based interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montreal, Quebec, Canada, 2006), CHI '06, ACM Press, pp. 653–662.

[32] Brewster, S., Wright, P. C., and Edwards, A. D. N. Experimentally Derived Guidelines for the Creation of Earcons. In *Adjunct Proceedings of HCI* (Huddersfield, UK, 1995), HCI '95, pp. 5–7.

[33] Brookmeyer, Gray, S., and Kawas, C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *American Journal of Public Health 88*, 9 (1998), 1337–1342.

[34] Brown, L. M., Brewster, S., and Purchase, H. C. A First Investigation into the Effectiveness of Tactons. In *First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems* (Pisa, Italy, 2005), IEEE Computer Society, pp. 167–176.

[35] Brown, L. M., Brewster, S., Ramloll, R., Burton, M., and Riedel, B. Design Guidelines for Audio Presentation of Graphs and Tables. In *International Conference on Auditory Display* (Boston, Massachusetts, USA, 2003), no. July in ICAD '03, pp. 284–287.

[36] Bruck, D., and Brennan, P. Recognition of Fire Cues During Sleep. In *Proceedings of the Second International Symposium on Human Behaviour in Fire* (London, UK, 2001), Interscience Communications, pp. 241–252.

[37] Brudzinski, M., Ratwani, R. M., and Trafton, J. G. Goal and spatial memory following interruption. *Proceedings of the Eighth* (2007).

[38] Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., and Channon, S. The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society 12*, 2 (Mar. 2006), 194–209.

[39] Burmistrov, I., and Leonova, A. Do Interrupted Users Work Faster or Slower? The Micro-analysis of Computerized Text Editing Task. In *Human-Computer Interaction: Theory and Practice - Proceedings of HCI International* (2003), vol. 1, Lawrence Erlbaum Associates, pp. 621–625.

[40] Cades, D., Trafton, J. G., and Boehm-Davis, D. A. Mitigating Disruptions: Can Resuming an Interrupted Task Be Trained? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*, 3 (2006), 368–371.

[41] Caprani, N., Greaney, J., and Porter, N. A Review of Memory Aid

Devices for an Ageing Population. *PsychNology Journal 4*, 3 (2006), 205–243.

[42] Carton, A. M., and Aiello, J. R. Control and Anticipation of Social Interruptions: Reduced Stress and Improved Task Performance. *Journal of Applied Social Psychology 39*, 1 (Jan. 2009), 169–185.

[43] Cellier, J.-M., and Eyrolle, H. Interference between switched tasks. *Ergonomics 35*, 1 (Jan. 1992), 25–36.

[44] Chappell, N. L., Dlitt, B. H., Hollander, M. J., Miller, J. A., and McWilliam, C. Comparative Costs of Home Care and Residential Care. *The Gerontologist 44*, 3 (June 2004), 389–400.

[45] Clifford, J. D., and Altmann, E. M. Managing Multiple Tasks : Reducing the Resumption Time of the Primary Task. *International Journal of Human-Computer Studies* (2003), 2003–2003.

[46] Cohen, S. Aftereffects of stress on human performance and social behavior: A review of research and theory. *Psychological Bulletin 88*, 1 (1980), 82–108.

[47] Cousineau, D. Confidence Intervalts in Within-Subject Designs: A Simple Solution to Loftus and Masson's Method. *Tutorial in Quantitative Methods for Psychology 1*, 1 (2005), 4–45.

[48] Crabtree, A., and Rodden, T. Domestic Routines and Design for the Home. *Computer Supported Cooperative Work 13*, 2 (2004), 191–220.

[49] Crabtree, A., Rodden, T., and Hemmings, T. Supporting Communication within Domestic Settings. In *Proceedings of the 2003 Home Oriented Informatics and Telematics Conference* (Irvine, California, USA, 2003).

[50] Crabtree, A., Rodden, T., Hemmings, T., and Benford, S. Finding a Place for UbiComp in the Home. *UbiComp 2003: Ubiquitous Computing 2864* (2003), 208–226.

[51] Cutrell, E., Czerwinski, M., and Horvitz, E. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Human-Computer Interaction 2001* (Tokyo, Japan, 2001), Interact '01, IOS Press, pp. 263–269.

[52] Danilov, Y., and Tyler, M. Brainport: An Alternative Input to the Brain. *Journal of Integrative Neuroscience 4*, 4 (2005), 537–550.

[53] Demirbilek, O., and Demirkan, H. Universal product design involving elderly users: a participatory design model. *Applied Ergonomics 35*, 4 (July 2004), 361–370.

[54] Dey, A. K., and Abowd, G. D. CybreMinder : A Context-Aware System for Supporting Reminders. In *Proceedings of the Second International Symposium on Handheld and Ubiquitous Computing* (Bristol, UK, 2000), vol. 1927 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 172–186.

[55] Dowdall, A., and Perry, M. The Millennium Home: Domestic Technology to Support Independent-living Older People. In *Proceedings of the 1st Equator IRC Workshop on Ubiquitous Computing in Domestic Environments* (Nottingham, UK, 2001), no. September 2001, pp. 1–15.

[56] Eccleston, C., and Crombez, G. Pain demands attention: a cognitive-affective model of the interruptive function of pain. *Psychological Bulletin 125*, 3 (May 1999), 356–366.

[57] Edwards, W. K., and Grinter, R. E. At Home with Ubiquitous Computing: Seven Challenges. In *UbiComp 2001: Ubiquitous Computing* (Atlanta Georgia, Oct. 2001), vol. 2201 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 256 – 272.

[58] Edworthy, J., Loxley, S., and Dennis, I. Improving Auditory Warning Design: Relationship between Warning Sound Parameters and Perceived Urgency. *Human Factors: The Journal of the Human Factors and Ergonomics 33*, 2 (1991), 205–231.

[59] Emery, V. K., Edwards, P. J., Jacko, J. A., Moloney, K. P., Barnard, L., Kongnakorn, T., Sainfort, F., and Scott, I. U. Toward achieving universal usability for older adults through multimodal feedback. In *Proceedings of the 2003 conference on Universal Usability* (Vancouver, British Columbia, Canada, 2003), CUU '03, ACM Press, pp. 46–53.

[60] Equality and Human Rights Commision. Close to Home: An inquiry into older people and human rights in home care., 2011.

[61] Espinoza, F., De Roure, D., Hamfors, O., Hinz, L., Holmberg, J., Jansson, C.-G., Jennings, N., Luck, M., Lönnqvist, P., Ramchurn, G., Sandin, A., Thomson, M., and Bylund, M. Intrusiveness Management for Focused, Efficient, and Enjoyable Activities. *The Disappearing Computer: Interaction Design, System Infrastructures and Applications for Smart Environments 4500* (2007), 143–157.

[62] Fickas, S. Clinical Requirements Engineering. In *Proceedings of the 27th international conference on Software engineering* (St. Louis, MO, USA, 2005), ICSE '05, ACM Press, pp. 28–34.

[63] Field, A. *Discovering Statistics using SPSS*, 3rd ed. SAGE Publications, 2009.

[64] Field, A., Miles, J., and Field, Z. *Discovering Stastistics using R*, 1st ed. SAGE Publications, 2012.

[65] Fleming, N. D., and Mills, C. Not Another Inventory, Rather a Catalyst for Reflection. *To Improve the Academy 11* (1992), 137–155.

[66] Galili, T. Post hoc Analysis for FriedmanâĂŹs Test, 2010.

[67] Gaver, W. W. Synthesizing auditory icons. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands, 1993), CHI '93, ACM Press, pp. 228–235.

[68] Gaver, W. W., and Gore, K. Auditory interfaces. In *Handbook of Human-Computer Interaction*, M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., 2nd ed. Elsevier Science, 1997, pp. 1003–1041.

[69] Gellatly, A., Jones, S., and Best, A. The Development of Skill at Concentration. *Australian Journal of Psychology 40*, 1 (1988), 1–10.

[70] Gibbs, W. W. Considerate Computing. *Scientific American 292*, 1 (Jan. 2005), 54–61.

[71] Gillie, T., and Broadbent, D. What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research 50*, 4 (Apr. 1989), 243–250.

[72] Giraldo, C., Helal, S., and Mann, W. C. mPCA âĂŞ A Mobile Patient Care-Giving Assistant for Alzheimer Patients. In *First International Workshop on Ubiquitous Computing for Cognitive Aids* (Göteborg, Sweden, 2002), UbiCog '02.

[73] Grasso, M. A., Ebert, D. S., and Finin, T. W. The integrality of speech in multimodal interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI) 5*, 4 (1998), 303.

[74] Groot, Y. C. T., Wilson, B. A., Evans, J., and Watson, P. Prospective memory functioning in people with and without brain injury. *Journal of the International Neuropsychological Society 8*, 5 (July 2002), 645–54.

[75] Grundgeiger, T., and Sanderson, P. Interruptions in Healthcare: Theoretical Views. *International Journal of Medical Informatics 78*, 5 (May 2009), 293–307.

[76] Hall, B., and Hursch, D. An Evaluation of the Effects of a Time Management Training Program on Work Efficiency. *Journal of Organizational Behavior Management 3*, 4 (Sept. 1982), 73–96.

[77] Halliday James. i-Cue Dynamic Support Environment, 2009.

[78] Hart, S. G. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*, 9 (2006), 904–908.

[79] Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload 52* (1988), 139–183.

[80] Hasofer, A., and Bruck, D. Statistical analysis of response to fire cues. *Fire Safety Journal 39*, 8 (Nov. 2004), 663–688.

[81] Hein, A., Nee, O., Willemsen, D., Scheffold, T., Dogac, A., Laleci, G. B., and Rothenfelde, S.-k. B. SAPHIRE- Intelligent Healthcare Monitoring based on Semantic Interoperability Platform- The Homecare Scenario-. In *1st European COnference on eHealth* (Fribourg, Switzerland, 2006), ECEH âĂŹ06, pp. 191–202.

[82] Helal, S., and Chen, C. The Gator Tech Smart House: Enabling Technologies and Lessons Learned. In *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology* (Singapore, 2009), i-CREATe '09, ACM Press, pp. 13:1—-13:4.

[83] Helal, S., Chen, C., Kim, E., Bose, R., and Lee, C. Toward an Ecosystem for Developing and Programming Assistive Environments. *Proceedings of the IEEE 100*, 8 (Aug. 2012), 2489–2504.

[84] Helal, S., Giraldo, C., Kaddoura, Y., El-Zabadani, H., and Mann, W. C. Smart phone based cognitive assistant. In *The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications* (Seattle, Washington, 2003), UbiHealth '03, Springer-Verlag.

[85] Helal, S., Mann, W. C., El-Zabadani, H., King, J., Kaddoura, Y., and Jansen, E. The Gator Tech Smart House: a programmable pervasive space. *Computer 38*, 3 (2005), 64–74.

[86] Hellier, E., Edworthy, J., Weedon, B., Walters, K., and Adams, A. The Perceived Urgency of Speech Warnings: Semantics versus Acoustics. *Human Factors: The Journal of the Human Factors and Ergonomics Society 44*, 1 (Jan. 2002), 1–17.

[87] Hodgetts, H. M., and Jones, D. M. Contextual Cues Aid Recovery From Interruption: The Role of Associative Activation. *Journal of Experimental Psychology 32*, 5 (2006), 1120–1132.

[88] Hodgetts, H. M., and Jones, D. M. Interruption of the Tower of London task: support for a goal-activation approach. *Journal of experimental psychology 135*, 1 (Feb. 2006), 103–115.

[89] Hodgetts, H. M., and Jones, D. M. Reminders, alerts and pop-ups: the cost of computer-initiated interruptions. In *Proceedings of the 12th international conference on Human-computer interaction: interaction design and usability* (Beijing, China, 2007), HCI '07, Springer-Verlag, pp. 818–826.

[90] Hoggan, E., and Brewster, S. Designing audio and tactile crossmodal icons for mobile devices. In *Proceedings of the 9th international conference on Multimodal interfaces* (Nagoya, Aichi, Japan, Nov. 2007), ICMI '07, ACM Press, pp. 162–169.

[91] Hoggan, E., Crossan, A., Brewster, S., and Kaaresoja, T. Audio or tactile feedback: which modality when? In *Proceedings of the 27th international conference on Human factors in computing systems* (Boston, MA, USA, 2009), CHI '09, ACM, pp. 2253–2256.

[92] Horvitz, E., and Apacible, J. Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces* (Vancouver, British Columbia, Canada, 2003), ICMI '03, ACM Press, pp. 20–27.

[93] Horvitz, E., Apacible, J., and Subramani, M. Balancing awareness and interruption: Investigation of notification deferral policies. In *Proceedings of the 10th International Conference on User Modeling* (Edinburgh, Scotland, UK, 2005), vol. 3538 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 433 – 437.

[94] Horvitz, E., Jacobs, A., and Hovel, D. Attention-Sensitive Alerting. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (Stockholm, Sweden, 1999), UAI '99, Morgan Kaufmann Publishers Inc., pp. 305–313.

[95] Horvitz, E., Kadie, C. M., Paek, T., and Hovel, D. Models of attention in computing and communication: from principles to applications. *Communications of the ACM 46*, 3 (2003), 52–59.

[96] Horvitz, E., Koch, P., and Apacible, J. BusyBody: creating and fielding personalized models of the cost of interruption. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work* (Chicago, Illinois, USA, 2004), vol. 6 of *CSCW '04*, ACM Press, pp. 507–510.

[97] Horvitz, E., Koch, P., Sarin, R., Apacible, J., and M. Bayesphone: Precomputation of context-sensitive policies for inquiry and action in mobile devices. In *Proceedings of the 10th International Conference on User Modeling*

(Edinburgh, Scotland, UK, 2005), UM 2005, Springer Berlin Heidelberg, pp. 251–260.

[98] Humble, J., Crabtree, A., Hemmings, T., Å kesson, K.-P., Koleva, B., Rodden, T., and Hansson, P. âĂIJPlaying with the BitsâĂİ User-Configuration of Ubiquitous Domestic Environments. In *Proceedings of the 5th International Conference on Ubiquitous Computing* (Seattle, WA, USA, 2003), vol. 2864 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 256 – 263.

[99] Intille, S. S. Designing a Home of the Future. *Pervasive Computing 1*, 2 (2002), 76–82.

[100] Intille, S. S. A New Research Challenge: Persuasive Technology to Motivate Healthy Aging. *IEEE transactions on information technology in biomedicine 8*, 3 (Sept. 2004), 235–7.

[101] Intille, S. S., Kukla, C., and Ma, X. Eliciting User Preferences Using Image-Based Experience Sampling and Reflection. In *CHI '02 extended abstracts on Human factors in computer systems* (New York, New York, USA, 2002), CHI '02, ACM Press, p. 738.

[102] Intille, S. S., Larson, K., Beaudin, J. S., Nawyn, J., Tapia, E. M., and Kaushik, P. A Living Laboratory for the Design and Evaluation of Ubiquitous Computing Technologies. In *CHI '05 extended abstracts on Human factors in computing systems* (Portland, Oregon, USA, 2005), CHI '05, ACM Press, pp. 1941–1944.

[103] Iqbal, S. T., and Bailey, B. P. Using Eye Gaze Patterns to Identify User Tasks. In *The Grace Hopper Celebration of Women in Computing* (2004), p. 6.

[104] Iqbal, S. T., and Bailey, B. P. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction 17*, 4 (Dec. 2010), 15:1–15:28.

[105] Iqbal, S. T., and Horvitz, E. Disruption and Recovery of Computing Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, 2007), CHI '07, ACM Press, pp. 677–686.

[106] ISO Technical Comitee 145 (Graphical Symbols). ISO 7001:2007 Graphical Symbols – Public Information Symbols, 2007.

[107] ISO Technical Comitee 145 (Graphical Symbols). ISO 7000:2012 Graphical symbols for use on equipment – Registered symbols, 2012.

[108] Jacko, J. A., Scott, I. U., Sainfort, F., Barnard, L., Edwards, P. J.,

Emery, V. K., Kongnakorn, T., Moloney, K. P., and Zorich, B. S. Older adults and visual impairment: what do exposure times and accuracy tell us about performance gains associated with multimodal feedback? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA, 2003), no. 5 in CHI '03, ACM Press, pp. 33–40.

[109] Joint Improvement Team and Association of Directors of Social Work. Analysis of Home Care National Statistical Returns and Local Authority Survey. Tech. Rep. February, Joint Improvement Team and Association of Directors of Social Work, 2009.

[110] Jones, B. R., Benko, H., Ofek, E., and Wilson, A. D. IllumiRoom : Peripheral Projected Illusions for Interactive Experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France, 2013), CHI '13, ACM Press, pp. 869–878.

[111] Jorge, J. A. Adaptive tools for the elderly. In *Proceedings of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing providing for the elderly* (Alcácer do Sal, Portugal, 2001), WUAUC '01, ACM Press, pp. 66–70.

[112] Ju, W., and Leifer, L. The Design of Implicit Interactions: Making Interactive Systems Less Obnoxious. *Design Issues 24*, 3 (July 2008), 72–84.

[113] Kane, R. A. Long-Term Care and a Good Quality of Life: Bringing Them Closer Together. *Gerontologist 41*, 3 (2001), 293–304.

[114] Kaneda, H., Maeshima, K., Goto, N., Kobayakawa, T., Ayabe-Kanamura, S., and Saito, S. Decline in taste and odor discrimination abilities with age, and relationship between gustation and olfaction. *Chemical Senses 25* (June 2000), 331–337.

[115] Kapitsa, M., and Blinnikova, I. Task performance under influence of interruptions. *Life and Behavioural Science 355* (2003), 323–329.

[116] Keoh, S. L., Dulay, N., Lupu, E., Twidle, K., Schaeffer-Filho, A. E., Sloman, M., Heeps, S., Strowes, S., and Sventek, J. Self-Managed Cell: A Middleware for Managing Body-Sensor Networks. In *Fourth Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services* (London, UK, 2007), MobiQuitous 2007, IEEE Computer Society, pp. 1–5.

[117] Khan, P., and Lenk, K. Principles of Typography for User Interface Design. *Interactions 5*, 6 (1998), 15–29.

[118] Kidd, C. D., Orr, R., Abowd, G. D., Atkeson, C. G., Essa, I. A., MacIntyre, B., Mynatt, E. D., Starner, T. E., and Newstetter, W. The Aware Home: A Living Laboratory for Ubiquitous Computing Research.

In *Proceedings of the Second International Workshop on Cooperative Buildings. Integrating Information, Organizations, and Architecture - CoBuildâĂŹ99* (Pittsburgh, PA, USA, 1999), vol. 1670 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 191–198.

[119] Kientz, J. A., Patel, S., Jones, B., Price, E., Mynatt, E. D., and Abowd, G. D. The Georgia Tech Aware Home. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (Florence, Italy, 2008), CHI EA '08, ACM Press, pp. 3675–3680.

[120] Kim, S., and James, T. W. Enhanced effectiveness in visuo-haptic object-selective brain regions with increasing stimulus salience. *Human Brain Mapping 31*, 5 (May 2010), 678–693.

[121] King, J., Bose, R., Yang, H.-I., Pickles, S., and Helal, S. Atlas: A Service-Oriented Sensor Platform: Hardware and Middleware to Enable Programmable Pervasive Spaces. In *Proceedings of the 31st IEEE Conference on Local Computer Networks* (Tampa, Florida, USA, 2006), IEEE Computer Society, pp. 630–638.

[122] Larson, K. The Home of the Future. *Architecture and Urbanism 10*, 361 (2000), 60–67.

[123] Latorella, K. A. Effects of Modality on Interrupted Flight Deck Performance: Implications for Data Link. Tech. rep., NASA Langley, 1998.

[124] Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y.-F., Field, A. S., and Stein, B. E. Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience 14*, 3 (Apr. 2002), 420–429.

[125] Lee, D., Helal, S., Anton, S., De Deugd, S., and Smith, A. Participatory and Persuasive Telehealth. *Gerontology 58*, 3 (Jan. 2011), 269–281.

[126] Lee, D., Helal, S., and Johnson, B. D. An Action-Based Behavior Model for Persuasive Telehealth. In *Proceedings of the 8th International Conference on Smart Homes and Health Telematics* (Seoul, Korea, 2010), vol. 6159 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 121–129.

[127] Lee, D., Yamazaki, T., and Helal, S. Robotic Companions for Smart Space Interactions. *Pervasive Computing 8*, 2 (2009), 78–84.

[128] Lee, J. D., Hoffman, J. D., and Hayes, E. Collision warning design to mitigate driver distraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria, 2004), vol. 6 of *CHI '04*, ACM Press, pp. 65–72.

[129] Levine, B., Stuss, D. T., Winocur, G., Binns, M. A., Fahy, L., Mandic, M., Bridges, K., and Robertson, I. H. Cognitive rehabilitation in the elderly: effects on strategic behavior in relation to goal management. *Journal of the International Neuropsychological Society 13*, 1 (Jan. 2007), 143–152.

[130] Leyvand, T., Meekhof, C., Wei, Y.-C., Sun, J., and Guo, B. Kinect Identity: Technology and Experience. *Computer 44*, 4 (Apr. 2011), 94–96.

[131] Living Made Easy. Telecare and Alarms. http://www.livingmadeeasy.org.uk/telecare%20and%20alarms/, 2013.

[132] Lupu, E., Dulay, N., Sloman, M., Sventek, J., Heeps, S., Strowes, S., Twidle, K., Keoh, S. L., and Schaeffer-Filho, A. E. AMUSE: autonomic management of ubiquitous e-Health systems. *Concurrency and Computation: Practice & Experience 20*, 3 (2008), 277–295.

[133] Manly, T., Hawkins, K., Evans, J., Woldt, K., and Robertson, I. H. Rehabilitation of executive function: facilitation of effective goal management on complex tasks using periodic auditory alerts. *Neuropsychologia 40*, 3 (Jan. 2002), 271–281.

[134] Mark, G., Gudith, D., and Klocke, U. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, 2008), CHI '08, ACM Press, pp. 107–110.

[135] Maternaghan, C. The Homer Home Automation System. *Technical Report*, CSM-187 (2010), 1–31.

[136] McAdam, C., and Brewster, S. Distal tactile feedback for text entry on tabletop computers. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (Cambridge, United Kingdom, 2009), BCS-HCI '09, British Computer Society, pp. 504–511.

[137] McBryan, T., and Gray, P. A Model-Based Approach to Supporting Configuration in Ubiquitous Systems. In *Interactive Systems. Design, Specification, and Verification* (Ontario, Canada, 2008), T. C. Graham and P. Palanque, Eds., Lecture notes in Computer Science, Springer-Verlag, pp. 167–180.

[138] McBryan, T., McGee-Lennon, M., and Gray, P. An Integrated Approach to Supporting Interaction Evolution in Home Care Systems. In *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments* (Athens, Greece, 2008), PETRA '08, ACM, pp. 45:1–45:8.

[139] McCrickard, D. S., Catrambone, R., Chewar, C., and Stasko, J. T. Establishing tradeoffs that leverage attention for utility: empirically evaluating information display in notification systems. *International Journal of Human-*

Computer Studies 58, 5 (2003), 547–582.

[140] McCrickard, D. S., and Chewar, C. M. Attuning notification design to user goals and attention costs. *Communications of the ACM 46*, 3 (2003), 67–72.

[141] McFarlane, D. C., and Latorella, K. A. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-Computer Interaction 17*, 1 (Mar. 2002), 1–61.

[142] McGee-Lennon, M. Requirements Engineering for Home Care Technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, 2008), CHI '08, ACM Press, pp. 1439–1442.

[143] McGee-Lennon, M., and Brewster, S. Reminders that make sense: Designing multimodal notifications for the home. In *Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare* (2011), PervasiveHealth, pp. 495–501.

[144] McGee-Lennon, M., Nigay, L., and Gray, P. The challenges of engineering multimodal interaction. *Journal on Multimodal User Interfaces 3*, 3 (Mar. 2010), 155–156.

[145] McGee-Lennon, M., Wolters, M., and Brewster, S. User-Centred Multimodal Reminders for Assistive Living. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada, 2011), CHI '11, ACM Press, pp. 2105–2114.

[146] McGee-Lennon, M., Wolters, M., and McBryan, T. Audio Reminders in the Home Environment. In *Proceedings of the 13th International Conference on Auditory Display* (Montreal, Canada, 2007), pp. 437–444.

[147] McGee-Lennon, M., Wolters, M., and McLachlan, R. Name that tune: musicons as reminders in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada, 2011), CHI '11, ACM Press, pp. 2803–2806.

[148] McGookin, D., and Brewster, S. Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Transactions on Applied Perception (TAP) 1*, 2 (2004), 130–155.

[149] McGurk, H., and MacDonald, J. Hearing lips and seeing voices. *Nature 264* (1976), 746 – 748.

[150] Mihailidis, A., Boger, J., Craig, T., and Hoey, J. The COACH prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics 8*, 28 (2008).

[151] Miller, G. a. The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychological Review 101*, 2 (Apr. 1955), 343–352.

[152] Miyata, Y., and Norman, D. Psychological issues in support of multiple activities. In *User Centered Systems Design: New Perspectives on Human-Computer Interaction*. CRC Press, 1986, ch. 13, pp. 265–284.

[153] Monk, C. A. The effect of frequent versus infrequent interruptions on primary task resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 48*, 3 (2004), 295–299.

[154] Monk, C. A., Boehm-Davis, D. A., and Trafton, J. G. The attentional costs of interrupting task performance at various stages. *Proceedings of the Human Factors and Ergonomics Society 46*, 22 (2002), 1824–1828.

[155] Monk, C. A., Boehm-Davis, D. A., and Trafton, J. G. Very brief interruptions result in resumption cost. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (2004), CogSci 2004, Citeseer, p. 1606.

[156] Morgan, M., Martin, C., McGee-Lennon, M., Clark, J., Hine, N., Wolters, M., and Arnott, J. Requirements gathering with diverse user groups and stakeholders. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (Florence, Italy, 2008), CHI EA '08, ACM Press, pp. 2597–2600.

[157] Morris, M. E., Intille, S. S., and Beaudin, J. S. Embedded Assessment: Overcoming Barriers to Early Detection with Pervasive Computing. In *Proceedings of the Third International Conference on Pervasive Computing: PERVASIVE* (Munich, Germany, 2005), vol. 3468 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 333–346.

[158] Mykityshyn, A. L., Fisk, A. D., and Rogers, W. A. Learning to Use a Home Medical Device: Mediating Age-Related Differences with Training. *Human Factors: The Journal of the Human Factors and Ergonomics Society 44*, 3 (2002), 354–364.

[159] Mynatt, E. D., Essa, I. A., and Rogers, W. A. Increasing the Opportunities for Aging in Place. In *Proceedings on the 2000 conference on Universal Usability* (Arlington, Virginia, USA, 2000), CUU '00, ACM Press, pp. 65–71.

[160] Mynatt, E. D., Melenhorst, A.-S., Fisk, A. D., and Rogers, W. A. Aware technologies for aging in place: Understanding user needs and attitudes. *Pervasive Computing 3*, 2 (2004), 36–41.

[161] Mynatt, E. D., Rowan, J., Craighill, S., and Jacobs, A. Digital Family Portraits: Supporting Peace of Mind for Extended Family Members. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA, 2001), CHI '01, ACM Press, pp. 333–340.

[162] Naumann, A. B., Wechsung, I., and Hurtienne, J. Multimodal interaction: A suitable strategy for including older users? *Interacting with Computers 22*, 6 (Nov. 2010), 465–474.

[163] Nawyn, J., Intille, S. S., and Larson, K. Embedding Behavior Modification Strategies into a Consumer Electronic Device: A Case Study. In *Proceedings of the 8th Internation Conference on Ubiquitous Computing* (Orange County, CA, USA, 2006), vol. 4206 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 297–314.

[164] Nikolic, M. I., and Sarter, N. B. Multimodal information presentation in support of timesharing and effective interruption management. In *Proceedings of the 20th Digital Avionics Systems Conference* (Daytona Beach, Florida, USA, 2001), vol. 1 of *20th DASC*, IEEE Computer Society, pp. 5D2/1–5D2/8.

[165] O'Brien, J., Rodden, T., Rouncefield, M., and Hughes, J. At home with the technology: an ethnographic study of a set-top-box trial. *ACM Transactions on Computer-Human Interaction 6*, 3 (1999), 282–308.

[166] Office for National Statistics UK. 2008-based National Population Projections: Frequently Asked Questions, 2008.

[167] Office for National Statistics UK. National Population Projections, 2008.

[168] Oviatt, S., and Cohen, P. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM 43*, 3 (Mar. 2000), 45–53.

[169] Park, D. Ageing and memory: mechanisms underlying age differences in performance. *Australasian Journal on Ageing 17* (1998), 69–72.

[170] Pashler, H., McDaniel, M., Rohrer, D., and Bjork, R. Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest 9*, 3 (2008), 105–119.

[171] Perry, M., Dowdall, A., Lines, L., and Hone, K. Multimodal and Ubiquitous Computing Systems: Supporting Independent-Living Older Users. *IEEE Transactions on Information Technology in Biomedicine 8*, 3 (2004), 258–270.

[172] Polli, A. Atmospherics/Weather Works: A Spatialized Meteorological Data

Sonification Project. *Leonardo 38*, 1 (2005), 31–36.

[173] Radvansky, G. A., Krawietz, S. A., and Tamplin, A. K. Walking Through Doorways Causes Forgetting: Further Explorations. *The Quarterly Journal of Experimental Psychology 64*, 8 (2011), 1632–1645.

[174] Reitter, D., Panttaja, E. M., and Cummins, F. UI on the fly: Generating a multimodal user interface. In *Proceedings of HLT-NAACL 2004: Short Papers* (Boston, Massachusetts, 2004), HLT-NAACL-Short '04, Association for Computational Linguistics, pp. 45–48.

[175] Ren, D., Zhou, H., and Fu, X. A Deeper Look at Gender Difference in Multitasking: Gender-Specific Mechanism of Cognitive Control. In *2009 Fifth International Conference on Natural Computation* (Tianjin, 2009), ICNC '09, IEEE Computer Society, pp. 13–17.

[176] Rogers, W. A., and Mynatt, E. D. How can technology contribute to the quality of life of older adults? In *The technology of humanity: Can technology contribute to the quality of life?* 2003, pp. 22–30.

[177] Rosenthal, R. *Meta-Analytic Procedures for Social Research*, 2nd ed. SAGE Publications, 1991.

[178] Russo, J., Sukojo, A., Helal, S., Davenport, R., and Mann, W. C. SmartWave âĂŞ Intelligent Meal Preparation System to Help Older People Live Independently. In *Proceedings of the Second International Conference on Smart homes and health Telematics* (Singapore, 2004), vol. 14 of *ICOST '04*, pp. 122–135.

[179] Rypma, B., and D'Esposito, M. Isolating the neural mechanisms of age-related changes in human working memory. *Nature Neuroscience 3*, 5 (May 2000), 509–515.

[180] Sanders, G. S., and Baron, R. S. The motivating effects of distraction on task performance. *Journal of Personality and Social Psychology 32*, 6 (1975), 956 – 963.

[181] Schumann-Hengsteler, R. Children's and adults' visuospatial memory: The game concentration. *Journal of Genetic Psychology 157*, 1 (1996), 77–92.

[182] Scott, S., Mercier, S., and Cummings, M. Assisting interruption recovery in supervisory control of multiple UAVs. *Human Factors and*, Figure 1 (2006), 699–703.

[183] Shallice, T., and Burgess, P. W. Deficits in Strategy Application Following Frontal Lobe Damage in Man. *Brain 114*, 2 (Apr. 1991), 727–741.

[184] Sharp, J. G., Byrne, J., and Bowker, R. The Trouble with VAK. *Educational Futures 1*, 1 (2007), 89–97.

[185] Shimojo, S., and Shams, L. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology 11*, 4 (2001), 505–509.

[186] Siio, I., Rowan, J., and Mynatt, E. D. Peek-A-Drawer : Communication by Furniture. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA, 2002), CHI '02, ACM Press, pp. 582–583.

[187] Siklos, S., and Kerns, K. A. Assessing multitasking in children with ADHD using a modified Six Elements Test. *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists 19*, 3 (Apr. 2004), 347–61.

[188] Smith, G., Della Sala, S., Logie, R. H., and Maylor, E. A. Prospective and Retrospective Memory in Normal Ageing and Dementia: A Questionnaire Study. *Memory 8*, 5 (Sept. 2000), 311–321.

[189] Speier, C., Valacich, J. S., and Vessey, I. The effects of task interruption and information presentation on individual decision making. In *Proceedings of the eighteenth international conference on Information Systems* (Atlanta, Georgia, USA, 1997), ICIS '97, Association for Information Systems, pp. 21–36.

[190] Stanford, V. Using Pervasive Computing to Deliver Elder Care. *Pervasive Computing 1*, 1 (2002), 10–13.

[191] Stockman, T., Hind, G., and Frauenberger, C. Interactive Sonification of Spreadsheets. In *International Conference on Auditory Display* (Limerick, Ireland, 2005), ICAD '05, pp. 134–139.

[192] Sweeney, S., Kersel, D., Morris, R. G., Manly, T., and Evans, J. J. The sensitivity of a virtual reality task to planning and prospective memory impairments: group differences and the efficacy of periodic alerts on performance. *Neuropsychological rehabilitation 20*, 2 (Apr. 2010), 239–263.

[193] Taylor, A. S., Swan, L., Eardley, R., Sellen, A., Hodges, S., and Wood, K. R. Augmenting Refrigerator Magnets: Why Less is sometimes More. In *Nordic Conference on Human-Computer Interaction* (Oslo, Norway, 2006), NordiCHI '06, ACM, pp. 115–124.

[194] The National Care Homes Research and Development Forum. My Home Life: Quality of Life in Care Homes. Tech. rep., Help the Aged, 2007.

[195] Trafton, J. G., Altmann, E. M., Brock, D. P., and Mintz, F. E. Preparing to resume an interrupted task: effects of prospective goal encoding and

retrospective rehearsal. *International Journal of Human-Computer Studies 58*, 5 (May 2003), 583–603.

[196] Tran, Q., Calcaterra, G., and Mynatt, E. How an Older and a Younger Adult Adopted a Cooking Memory Aid. *Proceedings of Human-Computer Interaction International* (2005).

[197] Tran, Q. T., Calcaterra, G., and Mynatt, E. D. CookâĂŹs Collage. In *Home-Oriented Informatics and Telematics*, A. Sloane, Ed. Springer US, 2005, pp. 15–32.

[198] Tran, Q. T., and Mynatt, E. D. What Was I Cooking ? Towards Déjà Vu Displays of Everyday Memory, 2003.

[199] Tsujita, H., Tsukada, K., and Itiro, S. InPhase: evaluation of a communication system focused on "happy coincidences" of daily behaviors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA, 2010), CHI '10, ACM Press, pp. 2481–2490.

[200] Tunstall. Portfolio of Telehealthcare Solutions, 2012.

[201] UK Department of Health. National Service Framework for Older People, 2007.

[202] Vastenburg, M. H., Keyson, D. V., and de Ridder, H. Considerate home notification systems: A user study of acceptability of notifications in a living-room laboratory. *International Journal of Human-Computer Studies 67*, 9 (Sept. 2009), 814–826.

[203] Vastenburg, M. H., Keyson, D. V., and Ridder, H. Considerate home notification systems: a field study of acceptability of notifications in the home. *Personal and Ubiquitous Computing 12*, 8 (June 2007), 555–566.

[204] Vaucelle, C., Bonanni, L., and Ishii, H. Design of haptic interfaces for therapy. In *Conference on Human Factors in Computing Systems* (Boston, 2009).

[205] Walker, B. N., and Cothran, J. T. Sonification Sandbox: A graphical toolkit for auditory graphs. In *International Conference on Auditory Display* (Boston, Massachusetts, USA, 2003), no. July in ICAD '03, pp. 161–163.

[206] Wang, F., and Turner, K. J. Policy Conflicts in Home Care Systems. In *9th International Conference on Feature Interactions in Software and Communications Systems* (Amsterdam, 2007), IOS Press.

[207] Wang, F., and Turner, K. J. Towards personalised home care systems. In *Proceedings of the 1st international conference on PErvasive Technologies Related*

to Assistive Environments (Athens, Greece, 2008), PETRA '08, ACM Press, pp. 44:1—-44:7.

[208] Wang, F., Turner, K. J., Docherty, L., Kolberg, M., and Magill, E. H. *Services and Policies for Care At Home.* IEEE Computer Society, 2006.

[209] Warnock, D. A Subjective Evaluation of Multimodal Notifications. In *Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare* (Dublin, Ireland, 2011), Pervasive Health, IEEE Computer Society, pp. 461–468.

[210] Warnock, D. The Application of Multiple Modalities for Improved Home Care Reminders. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA, 2012), CHI EA '12, ACM Press, pp. 951–954.

[211] Warnock, D., McGee-Lennon, M., and Brewster, S. The Impact of Unwanted Multimodal Notifications. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (Alicante, Spain, 2011), ICMI '11, ACM Press, pp. 177–184.

[212] Warnock, D., McGee-Lennon, M., and Brewster, S. The Role of Modality in Notification Performance. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - INTERACT 2011* (Portugal, Spain, 2011), P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler, Eds., vol. 6947 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 572–588.

[213] Warnock, D., McGee-Lennon, M., and Brewster, S. Older Users, Multimodal Reminders and Assisted Living Technology. *Health Informatics Journal 18*, 3 (2012), 181–190.

[214] Warnock, D., McGee-Lennon, M., and Brewster, S. Dynamic Multimodal Reminders for Home Care Systems. In *British HCI* (2013).

[215] Warnock, D., McGee-Lennon, M., and Brewster, S. Multiple Notification Modalities and Older Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France, 2013), CHI '13, ACM Press, pp. 1091–1094.

[216] Weffers-Albu, A., de Waele, S., Hoogenstraaten, W., and Kwisthout, C. Immersive TV viewing with advanced Ambilight. In *2011 IEEE International Conference on Consumer Electronics* (2011), ICCE, IEEE Computer Society, pp. 753–754.

[217] Weiser, M. The computer for the 21st century. *Scientific American 265*, 3 (1991), 94–105.

[218] Weiser, M., and Brown, J. S. Designing calm technology. *PowerGrid Journal 1*, 1 (1996), 75–85.

[219] Wiener, J. M., Hanley, R. J., Clark, R., and Van Nostrand, J. F. Measuring the activities of daily living: comparisons across national surveys. *Journal of Gerontology 45*, 6 (Nov. 1990), 229–237.

[220] Wilcox, R. *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed. Elsevier Inc., 2012.

[221] Williamson, J. R., McGee-Lennon, M., Freeman, E., and Brewster, S. Designing a Smartpen Reminder System for Older Adults. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France, 2013), CHI EA '13, ACM Press, pp. 73–78.

[222] Wilson, B. A., Evans, J. J., Emslie, H., Alderman, N., and Burgess, P. W. The development of an ecologically valid test for assessing patients with a dysexecutive syndrome. *Neuropsychological Rehabilitation 8*, 3 (May 1998), 213–228.

[223] Wilson, G., Brewster, S., Halvey, M., and Hughes, S. Thermal Icons: Evaluating Structured Thermal Feedback for Mobile Interaction. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services* (San Francisco, California, USA, 2012), MobileHCI '12, ACM Press, pp. 309–312.

[224] Wolters, M., Campbell, P., DePlacido, C., Liddell, A., and Owens, D. The effect of hearing loss on the intelligibility of synthetic speech. In *Proceedings of the 16th IInternational Congress of Phonetic Sciences* (Saarbrücken, Germany, 2007), ICPhS '07, pp. 673–676.

[225] World Health Organisation (WHO). World Health Statistics 2012, 2012.

[226] Xiao, B., Lunsford, R., Coulston, R., Wesson, M., and Oviatt, S. Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In *Proceedings of the 5th international conference on Multimodal interfaces* (Vancouver, British Columbia, Canada, 2003), ICMI '03, ACM Press, pp. 265–272.

[227] Young, S. L. Increasing the Noticeability of Warnings: Effects of Pictorial, Color, Signal Icon and Border. *Human Factors and Ergonomics Society Annual Meeting Proceedings 35*, 9 (Jan. 1991), 580–584.

[228] Zajicek, M., and Morrissey, W. Multimodality and Interactional Differences in Older Adults. *Universal Access in the Information Society 2*, 2 (2003), 125–133.

[229] ZAJICEK, M., WALES, R., AND LEE, A. Speech interaction for older adults. *Universal Access in the Information Society 3*, 2 (June 2004), 122–130.

[230] ZEIGARNIK, B. On Finished and Unfinished Tasks. *Psychologische Forschung 9* (1927), 1–15.

Glossary

Auditory Icon An auditory icon is the aural equivalent of a pictogram; a sound that can be interpreted to discern its meaning.

Cards Matched (CM) Cards Matched is a metric used in Chapters 3 and 4. It is a simple measure of performance representing the number of cards matched in a game. It is expressed as a percentage.

Carer A person who takes on the responsibility of proving assistive care to another person. This is usually a paid carer or family member.

Cross-Modal Plasticity The ability of the brain to reorganise itself so that the resources from an impaired sensory system can be used to strengthen other sensory systems.

Earcon An abstract structured sound that is used to deliver information.

Effort (EF) One of the six 21-point scales used in the NASA-TLX survey. This was measured with the following question: "*How hard did you have to work (mentally and physically) to accomplish your level of performance?*" [79].

Frustration (FR) One of the six 21-point scales used in the NASA-TLX survey. This was measured with the following question: "*How irritated, stresses, and annoyed versus content, relaxed, and complacent did you feel during the task?*" [79].

Gerontology The study of ageing in humans.

Impairment In the context of this thesis an impairment is a measurable decline in physical, sensory or cognitive ability.

Interruption Lag The time taken to switch from task A to task B after receiving a notification that task B needs to be attended to. This time is used to take a 'mental snapshot' of the current state of task A.

Mental Demand (MD) One of the six 21-point scales used in the NASA-TLX survey. This was measured with the following question: "*How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?*" [79].

Modal Learning Preference (MLP) Modal learning theory holds that people will have a natural 'preferred' way of learning, *e.g.* by reading or by doing. The preferred way of learning is called the 'modal learning preference'. There are several schools of though on the taxonomy of modes that can be used for learning; the most popular are VAK (Visual, Audio and Kinaesthetic) and VARK (Visual, Audio, Reading/Writing and Kinaesthetic).

Musicon A very short snippet (1-3 seconds) of music designed to behave as an abstract notification.

NASA Task-Load Index (NASA-TLX) NASA-TLX is a workload assessment survey that is comprised of 6 components; mental demand, physical demand, temporal demand, performance, effort and frustration. The overall workload is given by the sum of the 6 components. NASA-TLX was originally developed by Hart & Staveland [79] and included several pairwise comparisons to 'weight' the 6 workload components. Hart [78] later (20 years later, actually) noted that researchers often dropped the weighting part of the survey, but found that this did not result in a loss of reliability. The NASA-TLX survey was administered in every study in this thesis, but the component weighting was not carried out for practical reasons (*i.e.* timekeeping).

Notification There are several ways to deliver information to people; the term 'notifications' is generally used to describe prompts intended to deliver a piece of discrete information. There is no implication that this information is necessarily useful, and a great deal of literature has focussed on the reduction of low-value notifications. There is also nothing in the literature that implies a specific delivery method. It is however implied that the information is discrete; systems which provide continuous information are usually called alarms. Reminders are generally considered to be a time-bound sub-type of notification.

Older Person There is no standard age at which a person becomes 'old'; however '65 years or over' is the age most commonly cited in research and reports. However, for the purposes of this thesis the phrase 'older person' should be interpreted to mean 'a person who has developed a natural age-related impairment'.

OSGI OSGI is an open standard allowing the modalurisation of programs into 'bundles', which can be connected and disconnected from a system at runtime. OSGI is managed by the OSGI Alliance.

Overall Performance (OP) One of the six 21-point scales used in the NASA-TLX survey. This was measured with the following question: "*How successful were you in performing the task? How satisfied were you with your performance?*" [79].

Peltier Sometimes known as a peltier pump or heat pump, a peltier is an electronic device that gets cold at one side and hot at the other when an electric current is passed through it. It effectively 'pumps' heat from one side to the other.

Physical Demand (PD) One of the six 21-point scales used in the NASA-TLX survey. This was measured with the following question: "*How much physical activity was required? Was the task easy or demanding, slack or strenuous?*" [79].

Pictogram A pictogram/pictograph is an image that conveys some kind of meaning. Pictograms were the first known forms of writing. Computer icons are a type of pictogram, and some literature treats the terms 'icon' and 'pictogram' as synonyms.

Plugwise Plugwise is a power-line-based home automation tool similar to X10. Plugwise is more advanced than X10, and provides the ability to measure the amount of power being drawn by an electrical device.

Prospective Memory Prospective memory is used for organisation; it is essentially 'remembering to remember'. There are two types; internally cued, where a person remembers to do something without assistance, and externally cued, where a person remembers something due to associations with some event or object. This type of memory degrades naturally with age [41, 188].

Reminder Reminders are a type of notification used to prompt a person to carry out a time-bound task, or to notify a person that they have failed or forgotten to carry out a time-bound task. Examples include notifying someone to take medicine at a set time, to remember keys when leaving the house, or to turn off a cooking appliance if it has been left on. As this thesis is concerned with reminders, the terms 'notification' and 'reminders' are sometimes used interchangeably.

Resumption Lag The time taken to resume an interrupted task (task A) after attending to a secondary task (task B). This time is used to rebuild the mental state that existed before task A was interrupted.

Senescence The condition or process of deterioration with age; loss of a cell's power of division and growth.
Source: Oxford English Dictionary, British Edition.

Somatosensory System The distributed and heterogeneous network of receptors responsible for the sensations of pain, pressure and temperature.

Sonification The representation of data trends through non-speech audio; *e.g.* the clicks of a Geiger counter or the beeps of a heart rate monitor.

Tacton A tacton is the name given to a message delivered through tactile interaction, usually in the form of a vibration against the skin.

Telecare Technology that can support safety and independence in the home by using a network of sensors to help automate emergency assistance. Note that this is the definition given by the Disabled Living Foundation, and that the terms 'telehealth' and 'telehealthcare' are sometimes used interchangeably.

Telehealth Products which generally provide the following functions: *(1)* to assist a patient in managing or monitoring their medical conditions and *(2)* to provide richer communication between the patient and medical professionals or carers. Note that this is the definition given by the Disabled Living Foundation, and that the terms 'telecare' and 'telehealthcare' are sometimes used interchangeably.

Temporal Demand (TD) One of the six 21-point scales used in the NASA-TLX survey. This was measured with the following question: "*How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?*" [79].

Ubiquitous Computing (UbiComp) Technology that is so inexpensive and commonplace that it fades into the background of everyday life. There is some controversy over this definition, which is explored in Section 2.1.3 (p. 11).

Workload (WL) Workload is the sum of the six components of the NASA-TLX assessment. Note that while the workload can be weighted, this is not required [78] and therefore weights were not applied during when calculating workload in this thesis.

X10 Standard X10 is an industry standard for communication between home automation hardware, such as light switches and electrical sockets. It is an open standard to ensure compatibility between X10 products made by different manufacturers.

Zeigarnik Effect Discovered by Bluma Zeigarnik in 1927, the Zeigarnik effect suggests that people naturally want to complete tasks and that an interrupted task will remain in tension, causing people to remember more about it.

Index