JEEVES: A BLOCKS-BASED APPROACH TO END-USER DEVELOPMENT OF EXPERIENCE SAMPLING APPS

Daniel John Rough

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

2018

Jeeves: A Blocks-Based Approach to End-User Development of Experience Sampling Apps

Daniel John Rough

University of St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

March 2018

Abstract

Professional programmers are significantly outnumbered by end-users of software, and cannot possibly predict the diverse and dynamic needs of user groups in advance. This thesis is concerned with the provision of an end-user development (EUD) approach, allowing end-users to independently create and modify their own software.

EUD activities are particularly applicable to the work practices of psychology researchers and clinicians, who are increasingly dependent on software for assessment of participants and patients, but must also depend on developers to realise their requirements. This thesis targets these professionals, with an EUD solution to creating assessment software.

The Experience Sampling Method (ESM) is one such means of assessment that takes place in participants' everyday lives. Through regular completion of subjective self-reports, participants provide rich detail of their ongoing physical and emotional well-being. However, lack of engagement with such studies remains a prevalent issue. This thesis investigates features for maximising engagement with experience sampling smartphone apps.

Such apps are becoming accepted as standard practice for remote assessment, but researchers are stifled by the complexity and cost of implementation. Moreover, existing EUD tools are insufficient for development of ESM apps that include engaging features. This thesis presents the development of Jeeves, an EUD tool with a blocks-based programming paradigm that empowers non-programmers to rapidly develop tailored, context-sensitive ESM apps.

The adoption of Jeeves is contingent on a number of factors, including its ease-of-use, real-world utility, and organisational conditions. Failure to incorporate the necessary functionality pertaining to these factors into Jeeves will lead to abandonment. This thesis is concerned with establishing the usability, utility, and external factors necessary for adoption of Jeeves. Further, Jeeves is evaluated with respect to these factors through a series of rigorous studies from a range of application domains.

Acknowledgements

Where do I begin? I'd better start with Aaron as he's helped out with this thesis now and again. Thank you to Aaron Quigley, my supervisor, for being a pat on the back, and a kick up the arse when necessary; for always making the time for me; for your contagious energy; for not wringing my neck when it was probably justified; and for your compassion and understanding through all my ups and downs. You're a star.

Thank you also to my second supervisor, David Harris-Birtill, for always having a smile; for finding the good in everything; for your boundless enthusiasm; and for having patience that warrants a Nobel Peace Prize.

Big thanks to everyone in the School of Computer Science for being lovely folk these past 8 years and making me feel at home. I won't name anyone specific because you're all special in your own way, and honestly I'm struggling to remember any of you off the top of my head.

Last, and most importantly, thank you to my family. This thesis is almost 80,000 words. I could double that and then some if I tried to thank them for everything they've done, but I think we can agree it's probably long enough already.

Thank you Mam and Dad. For always being there; for your endless love; for never giving up on me; for doing everything you can to help; for always trying to understand, even when I can't understand myself.

Thank you Katie for picking me up at my lowest points; for always being there for chats, for inspiring me with your determination in everything you do; for being bloody hilarious; for being the best sister I could ever ask for.

...

To everyone else, I hate to be rude but you probably could've been a bit more helpful. Buy me a dram and all is forgiven.

Funding

This work was supported by a University of St Andrews 600th Anniversary PhD Scholarship (School of Computer Science).

Candidate's Declarations

I, Daniel John Rough, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 79,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2013.

I received funding from an organisation or institution and have acknowledged the funder in the full text of my thesis.

Date   Signature of candidate

Supervisor's Declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date   Signature of supervisor

Permission for Publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Daniel John Rough, have obtained, or am in the process of obtaining, third-party copyright permissions that are required or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date Signature of candidate

Date Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Daniel John Rough, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date Signature of candidate

Date Signature of supervisor

To Granny, Granda, Nan and Granda John x

Research Publications

Some of the ideas presented in this thesis have already been published in the following papers:

D. Rough and A. Quigley, "An end-user interface for behaviour change intervention development," in *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pp. 377–378, ACM, 2014

D. Rough and A. Quigley, "Jeeves - a visual programming environment for mobile experience sampling," in *Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on*, pp. 121–129, IEEE, 2015. **[Chapter 6]**

D. Rough and A. Quigley, "Jeeves - an experience sampling study creation tool," *BCS Health Informatics Scotland (HIS)*, 2017. **[Chapter 6]**

D. Rough and A. Quigley, "Overcoming mental blocks: a blocks-based approach to experience sampling studies," in *Blocks and Beyond Workshop (B&B), 2017 IEEE*, pp. 45–48, IEEE, 2017. **[Chapter 5]**

D. Rough and A. Quigley, "TAPping into mental models with blocks," in *Blocks and Beyond Workshop (B&B), 2017 IEEE*, pp. 115–116, IEEE,2017. **[Chapter 5]**

D. Rough and A. Quigley, "End-user development in social psychology research: factors for adoption," in *Visual Languages and Human-Centric Computing (VL/HCC), 2018 IEEE Symposium on*. IEEE, 2018. **[Chapter 7]**

D. Rough and A. Quigley, "Towards end-user development for chronic disease management," in *Designing Technologies to Support Human Problem Solving*. 2018. **[Chapters 7 and 8]**

CONTENTS

Chapter One

Introduction

Software is everywhere. Today, in activities such as writing a letter, driving a car, or cooking, it is highly likely that we are software end-users. In a relatively short space of time, computers have evolved from enormous mainframes owned by large organisations, to personal desktop machines in a moderate percentage of households, to the present day, where computers are forever with us, on us and around us. While the number of professional programmers has grown significantly with this evolution, it is dwarfed by the number of software end-users. To satisfy the increasingly diverse and dynamic requirements of these user groups, programmers can no longer remain the gatekeepers to software modifications, and instead must ensure that applications are flexible and customisable to the populations that use them. In short, software must become *soft*.

1.1 Research Context

The definition of *soft* with respect to software is its adaptability to a variety of end-users, providing each with a tailored experience. Perhaps the most pertinent example of this is the smartphone. Our smartphones are increasingly ubiquitous, yet are far removed from the original vision of ubiquitous computing. In 1991, Mark Weiser imagined that our technologies would *"weave themselves into the fabric of everyday life until they are indistinguishable from it"* [1, p. 94]. While they are undeniably part of our everyday lives, they do not disappear, and are still inherently *personal* computers. We have a certain attachment to our smartphones as personal devices, and to some extent, they are our digital identities [2], tailored to our preferences and lifestyles.

Smartphones and similar devices are not just personal toys, and in practically every field of research and industry, smartphone apps are becoming the standard for maintaining engagement with, and collecting data from, research participants and customers. From influencing consumer purchases through location-aware advertising, to encouraging healthy exercise and dietary habits, it is increasingly likely that *"there's an app for that"*.

1.1.1 Smartphones in healthcare

"The use of mobile and wireless technologies to support the achievement of health objectives (mHealth) has the potential to transform the face of health service delivery across the globe" - World Health Organization, 2011

This quote from the WHO's report on mHealth innovations [3, p. 1] exemplifies how app-based solutions to health service issues are being taken seriously by global organisations. Smartphone apps are well-placed to transform healthcare. The variety and popularity of health-related apps available for download are testament to the fact that health consumers are interested in managing and monitoring their health outside of the clinic, with a growing body of evidence demonstrating the beneficial health outcomes from supporting patients to manage their health independently.

This growth has occurred in parallel with that of the prevalence of mobile health apps, giving rise to the practice of 'mHealth'. mHealth is defined by the World Health Organization as *"medical and public health practice supported by mobile devices"* [3, p. 6]. Indeed, at the time of writing, there are now over 325,000 health apps available [4], provoking concerns about the lack of evidence-based information in the majority of these apps, and the associated dangers of misinforming their users [5]. In order to maximise the utility and minimise potential harm of app-based healthcare, the NHS currently provides a library of evaluated, NHS-certified apps in which *"trusted digital tools for patients and the public to manage and improve their health"* are available for download[1]. This thesis focuses on mHealth apps that are intended for patient use, and their benefits. Although most apps are intended to be used independent of clinician involvement, other apps are intended to be used as a means of communication between clinician and patient.

1.1.2 Smartphones in research

"It would border on scientific malpractice if we were still giving paper-and- pencil questionnaires to a few hundred local college students, recruiting a few dozen people to participate in laboratory tasks, or running Internet studies for people just sitting at desks" [6, p. 222].

This quote from Geoffrey Miller's *Smartphone Psychology Manifesto* exemplifies researchers' interest in the potential to capture rich, informative behavioural data with smartphones. Miller advocates a re-evaluation of psychology research methods in the wake of new possibilities

[1] https://digital.nhs.uk/nhs-apps-library Accessed 17/03/18

afforded by smartphones. This section thus describes how smartphone apps have exposed psychological processes in ways previously impossible using traditional research methods.

While physiological data captured through smartphones can expose participants' physical health, smartphone apps are also employed to capture *psychological* information. With a growing number of mental disability diagnoses, the ability to understand thoughts and feelings of individuals and groups in different contexts has never been more valuable. Observing participants in their natural environments is the gold standard for capturing their everyday experiences with high ecological validity, free from bias inherent in self-report. However, the process of doing so is both burdensome for the researcher and intrusive for the participant. In lieu of this, smartphone sensors, including GPS chips and accelerometers, as well as data streams such as phone call frequency and social media interactions, can be used to provide a naturalistic insight into participants' lives. Miller describes how such sensors have enabled automatic collection of user information, free from the need for direct observation or manual self-report [6]. Adams et al. further elaborate on these capabilities in a recent review, describing how both physiological and behavioural characteristics of smartphone users could be passively captured [7].

Sensor-based inference is not always accurate nor appropriate for capturing psychological states. As a compromise, diary studies have participants complete timely self-reports in their everyday contexts, but these raise their own validity questions. For example, participants may not want to carry paper diaries around, they may not fill them in at the right times. Further, the laborious process of transcribing paper data can introduce its own human errors [8]. In light of this, smartphones are increasingly used for remote, repeated data collection. By employing smartphones in this way, psychological data can augment physiological measures, while minimising the burden on participants. This research method is known as *experience sampling*, and is a focal point of this thesis.

1.1.3 Experience sampling

The Experience Sampling Method (ESM) is described by Csikszentmihalyi and Larson as *"an attempt to provide a valid instrument to describe variations in self-reports of mental processes"* [9, p. 526]. A similar definition, in the context of behavioural medicine, is *Ecological Momentary Assessment* (EMA), given by Stone and Shiffman in 1994 [10]. Although ESM and EMA evolved separately within their respective fields, today they are often used interchangeably to represent any method involving the repeated, remote assessment of participants. Further, the term *Ambulatory Assessment* (AA) was used by Fahrenberg in 1996 to define methods that also encompasses physiological measurements such as blood pressure, gait analysis or galvanic skin response [11] in addition to subjective self-report. Although collection of sensor data is a relevant

Table 1.1: Terminology of remote, repeated data capture methods, with definitions from literature

Term	Field	Definition
Experience Sampling Method	Psychology	"A method in which recording of feelings and activities is done on-line at the moment, either at randomly selected moments or at predetermined times" [12, p. 157]
Ecological Momentary Assessment	Medicine	"Repeated sampling of subjects' current behaviors and experiences in real time, in subjects' natural environments" [13, p. 1]
Ambulatory Assessment	Both	"A wide range of assessment methods to study people in their natural environment, including self-report, observational, and biological/physiological/behavioral" [14, p. 151]

advantage, the focus of this thesis is primarily on applications that can be run without external hardware such as blood pressure monitors, for example. These definitions, and their fields of origin, are summarised in Table 1.1. For the purposes of this thesis, the term ESM will be used exclusively to refer to any application of the methodology for improving research or clinical practice.

ESM attempts to minimise recall bias and enhance *ecological validity* - the extent to which the findings of a research study are able to be generalised to real-world settings. Conceptually, participants in ESM studies are instructed to provide details on their current physical or mental state, at various times, as they go about their daily lives. This contrasts with typical laboratory-based assessments, where participants are asked to recall such details that may have occurred days before, or to perform short, rigorously controlled experiments that fail to capture the natural contexts of emotions or symptoms.

Preceding the advent of cheap personal computing, ESM studies were conducted with paper diaries that participants would fill in at various times during the day. These would often be accompanied by a basic signalling device that would randomly prompt participants to fill in a diary. Since then, smartphones have become increasingly ubiquitous across all ages and demographics, providing an opportunity to overcome many of the limitations of paper, by deploying ESM applications directly onto participants' smartphones. Using this method, the phone acts as both the signalling device and the medium for delivering surveys, through which participants can complete and send their survey results to researchers in real-time.

1.2 Problem Statement

While the benefits of ESM smartphone apps over paper methods are apparent, developing these apps still requires programming experience. Researchers and clinicians who would benefit

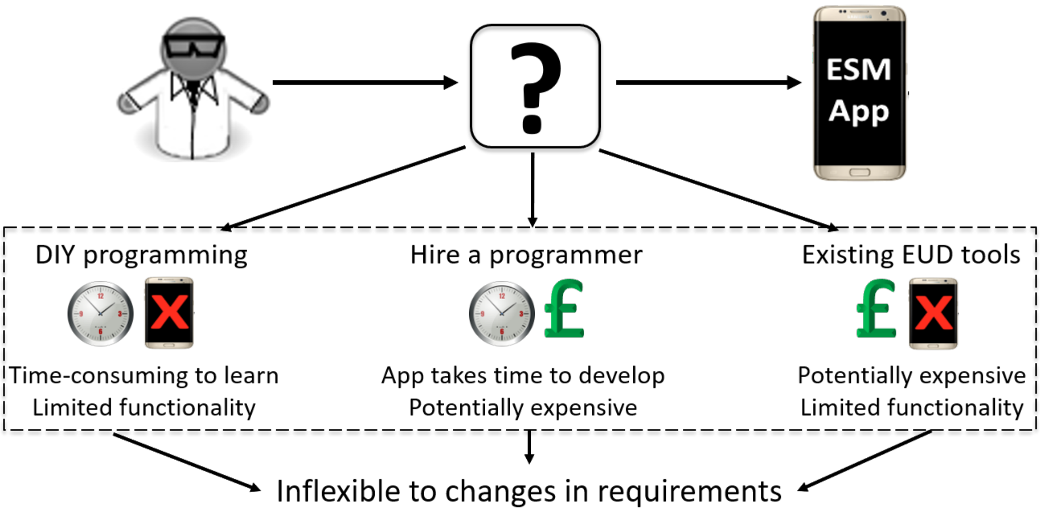


Figure 1.1: Time, cost, limited functionality, and inflexibility may inhibit ESM app creation

from ESM apps generally do not have this experience, nor the time and motivation to acquire it. Similarly, professional programmers are unlikely to have the domain-specific knowledge to understand researchers' requirements. Irrespective of the financial cost of professional development, further modifications to an app are then dependent on a programmer, and will incur additional time and cost to realise. Another alternative is to use commercially available tools that provide interfaces for creating simple ESM studies, a number of which are discussed in Chapter 4. In addition to their cost, particularly for medium to large-scale studies, all currently available commercial tools are still inflexible in terms of allowing a researcher to capitalise on the full potential of real-time, personalised, context-sensitive ESM apps.

Free and open-source tools are also available for use in the development of smartphone ESM apps, but as the review of recent ESM studies in Chapter 2 highlights, the majority of studies do not reference the use of such software. Indeed, with the lack of standardisation caused by research groups continuing to implement custom apps from scratch, this contributes to the difficulty in replicating their results, defined as a "replicability crisis" [15]. In the literature reviewed in Chapter 2, publications of studies focus on results and statistical analyses. While this is indisputably important, it comes at the expense of brief, generalised descriptions of study specifications. Although textual descriptions of these specifications could be verbose, their omission prevents accurate replication by other researchers in future. With inconsistencies in the implementation and reporting of ESM studies, researchers will encounter issues that others have already faced, and repeat mistakes in data collection and analysis that could have been avoided.

Chapter 3 presents a review of user-centred design studies from which design guidelines for smartphone ESM apps are derived, but despite the clear utility of such features, their availability in existing creation tools is highly limited, as is their use in recent ESM studies. In summary, as illustrated in Figure 1.1, it appears that the uptake of smartphone-based ESM in research and practice is currently hindered by difficulty in implementing studies, and the high cost and inflexibility of existing tools to facilitate this implementation. However, there is no hard evidence to confirm that this is the case, and it may be that custom implementations of basic ESM study specifications are satisfactory for researchers, alleviating the need for EUD solutions. As such, the work of this thesis is not simply tool development, but rather the investigation of reasons behind why other tools are not consistently adopted, and what would motivate researchers to do so. This is addressed through answering the research questions defined in the following section.

1.3 Research Questions and Objectives

The work in this thesis focuses on *end-user development* (EUD), a field of human-computer interaction research that investigates how end-users of software can be empowered to create and modify such software to satisfy their own diverse and dynamic requirements [16]. The underlying motivation is that, by allowing researchers to develop and deploy their ESM apps independent of a professional programmer, this will significantly lower the existing barriers to addressing research questions and healthcare issues that could benefit from such apps.

Towards understanding these existing barriers and how they could be overcome, this thesis describes the design, development and analysis of **Jeeves** (Java End-user Environment for Visual Experience Sampling) a desktop-based end-user development tool that allows researchers with no programming skills to create their own ESM apps, couple them to participants' smartphones, and receive results in real-time. By deriving design requirements from previous literature, implementing a prototype and evaluating it in terms of usefulness and usability, the following research question is addressed:

> Research Question
>
> ***What factors influence the adoption of technology for researchers and clinicians to develop experience sampling smartphone apps?***

The question is one of technology acceptance, narrowed in scope by the type of technology (an EUD tool) and the specific domain of ESM. However, much research in end-user development focuses on the implementation of a novel environment, and its subsequent usability evaluation. Although usability is critical for the adoption of end-user development technology, it is often vaguely defined, and dependent on the context of use. Moreover, it is a single factor that does not take into account whether the technology would actually be useful, nor whether it could be feasibly deployed in practice.

From this overarching question, four sub-questions have been derived that, while not mutually independent, can be separately investigated. These sub-questions are defined below, followed by the methods applied in this thesis to answer them.

Research Sub-Questions

- What difficulties do researchers and clinicians face that allowing them to develop experience sampling apps could alleviate?
- What are the necessary features of apps that facilitate remote assessment of research participants or clinical patients in their natural environments?
- How can the development effort be reduced to allow rapid creation of experience sampling apps without requiring programming experience?
- How can researchers and clinicians employ an experience sampling app end-user development tool in practice?

1.3.1 Research methods

The research methods undertaken to answer this question can be classified using the scheme developed by Kjeldskov and Graham for categorising research methods in Mobile HCI [17]. Five different research purposes are described as follows, with definitions taken from work by Tetteroo and Markopoulos [18]:

- **Understanding:** Research aimed at understanding the particulars of a phenomenon studied
- **Engineering:** Research aimed at the original development of a tool or technology
- **Re-engineering:** Research aimed at the engineering of modifications or extensions to an existing tool or technology
- **Evaluating:** Research aimed at the assessment, validation and assurance of tools, technology, models and frameworks
- **Describing:** Research aimed at describing the ideal properties of a system or situation

Each of these sub-questions are further broken down into concrete objectives, defined by their research *purpose* as described above, and the research *methods* applied.

1.3.2 What difficulties do researchers and clinicians face?

Objective 1 - Gather feedback from end-users with a potential interest in ESM studies, namely researchers in social psychology and practising clinicians. Through semi-structured interviews, gain insight into their current working practices, use of technology to support research and practice, and difficulties in doing so. This means of **understanding** will be accomplished through **survey research**.

Objective 2 - Observe the working practices of potential end-users of Jeeves, in order to determine any barriers related to usage context. This observation should be as naturalistic as possible, and is again a type of **understanding** research, achieved through a **field study**.

1.3.3 What are the necessary features of apps?

Objective 3 - Review recent publications that describe a study utilising experience sampling. Extract the different technologies used, sampling strategies, compliance rates and other study-specific factors where they have been appropriately described. In this objective, one goal is **understanding** how ESM studies are currently conducted, with a further goal of **describing** useful properties of ESM apps. As a literature review, it is classified as **basic research**.

Objective 4 - Review the current state-of-the-art in ESM app creation software. A source of such tools will be publications reviewed as part of the previous objective, but will also take into account commercial tools found via web search engines, and in particular, tools found from searching computer science literature databases. Again, as a literature review, this is **basic** research, with the purpose of **describing** the minimum requirements of a novel tool.

1.3.4 How can the development effort be reduced?

Objective 5 - Utilising interdisciplinary knowledge of EUD and ESM studies, both theoretical and practical, design and implement a prototype application (Jeeves) allowing non-programmers to create their own ESM smartphone apps. This is an objective with the purpose of **engineering**, achieved through the method of **applied research**.

Objective 6 - Using feedback from users with mixed experience of programming, iteratively develop Jeeves to maximise its universal usability. User feedback will take the form of quantitative data on task times and error rates, as well as qualitative data derived from interview feedback following task-based usability studies. This **evaluating** research will be conducted through **laboratory experiments**.

1.3.5 How can a development tool be used in practice?

Objective 7 - Following usability feedback gathered from the previous objective, conduct interviews with researchers and clinicians addressing the perceived usefulness of Jeeves in their domain of research or practice. Again, this is **evaluating** research, focused on the real-world utility of Jeeves, through **survey research** with potential end-users.

Objective 8 - To evaluate researchers' practical application of Jeeves, rather than simply their *perceived* utility, conduct **case studies** with researchers in which they apply Jeeves in addressing their own research questions. This is further **evaluating** research, demonstrating the utility *and* usability of Jeeves in its context of use.

1.4 Contributions

Through the interdisciplinary work described in this thesis, the following theoretical contributions are of value to human-computer interaction, as well as psychology research and clinical practice.

- A literature review of user-centred design studies and strategies for maximising ESM study utility, from which guidelines for future ESM tools are derived (Chapter 3)
- A series of user studies that demonstrate the usability of the Jeeves blocks-based programming paradigm, with implications for domain-specific EUD and blocks-based programming (Chapter 6)
- A qualitative analysis of interviews, observations and case studies involving researchers and clinicians, from which a set of requirements are derived for ESM app EUD, with implications for other novel technology deployed in such domains (Chapters 7 & 8)
- A set of design guidelines for the development of future EUD-ESM tools, consolidated from the results of analytical and empirical research (Chapter 9)

Further, the Jeeves platform represents a practical contribution, which can be utilised by non-programmers interested in running ESM studies. It consists of:

- The Jeeves blocks-based programming tool, allowing non-programmers to specify tailored, context-sensitive ESM apps, collect data and monitor compliance in real-time
- A dynamic Android ESM app (JeevesAndroid) that has been developed in parallel with the visual programming environment that can run any apps created with the Jeeves tool
- A framework that allows apps to be modified in real-time, and implicit and explicit self-report data to be instantly accessible by researchers

Table 1.2: Objectives addressed in each of the primary thesis chapters

Chapter	Objectives Addressed
2	O3 - Review recent ESM studies in psychology and medicine, compare and contrast with ESM studies in computer science
3/4	O4 - Review state-of-the-art EUD tools and address limitations
5	O5 - Implement Jeeves, JeevesAndroid and the overall framework
6	O6 - Conduct targeted, task-based usability evaluations of Jeeves
7	O1 - Understand working practices of researchers and clinicians O7 - Evaluate researchers' and clinicians' perceived utility of Jeeves
8	O2 - Observe working practices of potential clinician end-users O8 - Conduct case studies with psychology researchers using Jeeves

1.5 Thesis Overview

This thesis is organised into nine chapters. Table 1.2 describes where the objectives outlined in previous sections are addressed in each of these chapters. As can be seen from this table, the objectives are not addressed in order. Further, the working practices of researchers and clinicians, and associated requirements, were not addressed until after implementation of a fully-working prototype. Justification for this limitation is described in Chapter 9.

Chapter 2 provides detailed background information on ESM, explaining the various benefits and challenges of ESM, its use in research as well as clinical practice, and the evolving contribution of technology to ESM. Further, a review of recent ESM studies is presented, demonstrating the current challenges.

Chapter 3 reviews background literature on the use of smartphone apps in healthcare and psychology, deriving guidelines for future ESM tools from their stakeholders' perspectives. From this, a model is synthesised that represents these features as interactions between apps, researchers and participants.

Chapter 4 reviews background literature related to EUD. From this, a review of tools for ESM study creation is presented, analysing each tool with respect to the app features identified in the previous chapter, providing motivation for the implementation of Jeeves.

Chapter 5 presents design requirements of an EUD tool for ESM app creation, based on the literature reviewed in Chapters 3 and 4, and how the design of Jeeves satisfies these requirements. Design decisions of Jeeves are described in relation to established design frameworks and principles. Finally, a description of the full platform is given, including the JeevesAndroid app and overall architecture.

Chapter 6 describes three studies that were conducted to investigate the usability of Jeeves for non-programmers, and the iterative development improvements that were made as a result of feedback from these studies. Additionally, design guidelines for domain-specific visual programming environments are derived from this study feedback.

Chapter 7 details a qualitative analysis of interviews that assess the potential contribution of Jeeves to psychology research and clinical practice. These interviews discuss current working practices and associated difficulties of researchers and clinicians, and elicit their preliminary feedback on the Jeeves prototype. From these interviews, further factors addressing the utility of a tool for ESM app creation are derived.

Chapter 8 describes case studies that validate the efficacy of Jeeves for creating a study specification and running it with participants, with further feedback on real-world usability and utility. It also describes an observation session that was conducted in a local clinic to determine further factors for adoption in this domain.

Chapter 9, the conclusion, summarises the contribution of this thesis to the field of end-user development, as well as the interdisciplinary contribution to psychology and medicine. This conclusion also postulates future work planned for Jeeves and the limitations of this research.

CHAPTER TWO

EXPERIENCE SAMPLING

This chapter describes the Experience Sampling Method (ESM) in more detail, including the general methodology, its advantages and disadvantages, and relevant domain-specific concepts. ESM, and the closely related method of Ecological Momentary Assessment, have their roots in psychology and clinical research respectively. In human-computer interaction, the methodology has been borrowed for evaluating how users interact with technology in natural contexts, just as ethnographic methods have been borrowed from social science for the same purpose. As a contribution to the method, human-computer interaction researchers can develop interfaces that facilitate the creation of ESM smartphone apps - the overarching goal of this thesis.

This chapter defines ESM in more detail, describes its benefits and limitations, and how it has contributed to research in a variety of fields. Further, a review of recent ESM literature has been conducted, providing insights of relevance to this thesis.

2.1 Overview

Experience sampling studies involve asking participants to complete assessments on their current thoughts, feelings or context, as they go about their everyday lives. Indeed, the aforementioned *Ecological Momentary Assessment* summarises the defining features of the method in its title:

- **Ecological:** Data collection takes place *in situ*, as participants go about their everyday lives, as opposed to an artificial lab environment.
- **Momentary:** Participants are asked to report their state *in-the-moment*, rather than summarise over many hours, days or weeks.
- **Assessment:** Participants provide data through explicit self-report of their current states (although implicit data is also used as a secondary source, as described in Section 3.1).

Table 2.1: Summary of main benefits and challenges of experience sampling

Benefits	Challenges
Participants are assessed in their natural environments - **high ecological validity**	**Implementation can be difficult**, costly and burdensome for researchers and participants
Participants are asked about events close to their occurrence, **minimising recall bias**	Continuous assessment can influence variables of interest - **assessment reactivity**
Repeated assessment allows observation of **dynamic and contextual associations**	Repeated assessment places high burden on participants, provoking **non-compliance**

These features of ESM provide benefits over other qualitative research methods of data collection such as retrospective self-report, or face-to-face interviews. Nevertheless, these benefits also present challenges that require further research and interdisciplinary knowledge to overcome. This section will discuss both these benefits and challenges, and their implications for research, with a summary shown in Table 2.1.

2.1.1 Benefits

The benefits of ESM for capturing the intricacies of daily life as they occur have been widely acknowledged and described in various reviews on the methodology. This section summarises these benefits, which have justified its continued use in psychology and medicine, and its novel application in other fields. The following benefits are agnostic to the technology used, and are thus applicable to paper-based applications of ESM as well as mobile technology. (More specific advantages of smartphones for ESM are detailed further in Section 3.1.)

2.1.1.1 Ecological Validity

The validity of ESM self-reports are enhanced by both the temporal constraints on assessments, as well as the contexts in which these assessments take place. Having participants report on experiences, feelings, or other variables in their natural environment increases the ecological validity of assessments [19]. Conversely, participants in an unfamiliar laboratory or clinic are less likely to disclose sensitive information. In medicine, a relevant phenomenon is that of "white-coat hypertension", where patients experience raised blood pressure in direct response to the clinical environment.

The characteristics of a participant's context, including location, time, or presence of others, may be directly linked to the emergence of specific thoughts or behaviours [14]. Traditional assessments that take place at a lab or clinic fail to accurately capture the context of symptoms, whereas experience sampling measures can be used to acquire this context at the time of assessment. This is necessary for monitoring certain behaviours; for example, substance use is highly dependent on situational cues and social context [20].

2.1.1.2 Minimal Recall Bias

The *momentary* aspect of experience sampling alleviates participants from relying on recollection of past experiences. When asked to recall emotions or physical symptoms from a previous point in time, responses are inherently subjected to "recall bias". The passage of time can distort one's perception of a particular incident, and indeed one's current state can influence recollection of past state. Research has also shown that participants may employ a "peak-end rule", whereby summaries of experience over time are biased towards the most salient and recent experiences [21].

Experience sampling minimises this bias by reducing the time between an event of interest and the time at which this event is assessed. Further, summaries of experience are largely unnecessary, as only participants' immediate state is requested at each repeated assessment [13]. Studies have indeed identified the discrepancy between real-time data collection and retrospective self-report methods [22, 23], emphasising their importance for researchers in capturing less biased, contextually relevant data from participants.

2.1.1.3 Dynamic and Contextual Associations

The longitudinal aspect of experience sampling enables the collection of ecologically valid data for sustained periods of time and in a variety of contexts. In contrast, other assessment methods involving cross-sectional data collection limit possible analyses to differences between participants at specific contexts or instances of time. In capturing within-participant differences across time and contexts, ESM data offers a variety of analysis opportunities. Shiffman et al. characterise four such uses [13], which are explained and justified with additional examples.

Individual differences

Complementary to providing insight into within-person variations, experience sampling data can also be *aggregated* across time to provide detailed insight into individual differences. In such analyses, participants themselves are independent variables, and differences between participants at specific points in time have greater validity, as these differences are aggregated from multiple data points, as opposed to a single cross-sectional sample. Examples of questions that can be answered include whether men and women differ in how they experience a particular event, or how the daily emotional experiences of those with mental disorders differ from a control group.

Myin-Germeys et al. used experience sampling to distinguish levels of emotional reactivity between participants with three different psychological disorders [24]. Trull et al. similarly distinguished emotional instability between participants with borderline personality disorder (BPD) and those with depressive disorder [25]. While the emotional instability of the BPD participants in comparison to the depressed participants was known prior to the study, these results validated experience sampling for exposing such differences.

Describing history

While aggregation of ESM data is useful for obtaining accurate between-participant variation, a key benefit of experience sampling is that the resulting longitudinal data exposes *within*-participant differences. This is particularly important for assessing traits that can fluctuate rapidly over time, allowing researchers to study dynamic processes. Ebner-Priemer and Trull discuss this in their review of ESM in mood disorders, where studies were used to quantify the instability of individuals with BPD over a period of time [26].

Scollon et al. refer to within-participant analyses as *idiographic* research, as opposed to *nomothetic* research [12], which investigates group variations. In the clinical research domain, repeated, idiographic assessments are necessary to capture the dynamic factors of certain conditions. For example, instability cannot be observed in cross-sectional assessments of other conditions characterised by rapid fluctuations, such as psychosis and bipolar disorder.

Temporal sequences

Insights obtained from within-participant variations in idiographic research are largely independent of the sequence in which these variations occur. However, Shiffman et al. describe how the temporal sequence of states, in addition to their length and magnitude, can provide a further source of analysis. A number of studies have investigated the contextual factors that surround the occurrence of events in particular domains. For example, Haedt-Matt and Keel present a meta-analysis of studies using ESM to investigate the precursors and consequences of binge eating [27]. A study by Mitchell et al. found significant antecedents and consequences of smoking behaviour through experience sampling, with implications for clinical practice [28].

Contextual factors

Time is just one constituent of the context surrounding participants in their daily experiences. Further sources of context can be acquired in a number of ways - for example, they can be assessed through standard self-report measures, by explicitly asking participants where they are, whom they are with, or what they are doing. Shiffman et al. found significant correlations between smoking behaviour, location and presence of others by capturing these variables through self-report [20]. Additionally, certain contextual information can be captured through external devices. For example, Dunton et al. were able to investigate the contextual factors of physical activity by employing accelerometers [29]; Ilies et al. found associations between affect and cardiovascular functioning in a workplace environment with a wearable heart-rate monitor [30]. In recent years, contextual variables have also been captured by sensors embedded in smartphones, the capabilities and advantages of which have been revered for behavioural and clinical research [7, 31]. Further detail on sensor use in ESM is discussed in Section 3.1.

Treatment monitoring

Shiffman et al. do not explicitly mention treatment monitoring in their applications of experience sampling, instead focusing on its usage in research. However, such an application offers unique advantages in clinical practice. For example, frequent travel to appointments is time-consuming and costly, as well as physically difficult for patients with disabilities. As a potential solution, the remote monitoring of symptoms through ESM could reduce the number of routine face-to-face appointments that are required. This reciprocally benefits clinicians themselves, for whom time is saved through supporting patients' independence in monitoring their own health.

In one study, Wichers et al. reviewed the potential for ESM to assist in the treatment of patients with clinical depression, with quantified, visualised feedback on symptomatic states allowing for *"patients and clinicians alike to understand, modify and track the experiences that are currently subsumed under the 'black box' diagnosis of depression"* [32, p. 264]. A recent review by van Os et al. further discusses the advantages of smartphone technology to ESM in practice [33]. Examples include the regular assessment of symptoms and emotions to identify optimal doses of medication, and to uncover correlations between co-existing symptoms. ESM can even promote positive mental states, where the simple act of self-monitoring and receiving feedback can act as a form of behavioural intervention [34].

2.1.2 Challenges

Although ESM provides a number of benefits for research and practice, the nature of such an intensive data collection protocol innately has complications, which this section will discuss in more detail. While some of these disadvantages are agnostic of technology used, others have been mitigated in recent years through smartphone-based applications. Indeed, technological advancements that could further alleviate the following challenges are an active area of research.

2.1.2.1 Participant burden

The completion of assessments at various times of the day can interrupt the normal activities of participants' lives. Traditional paper diary studies are particularly burdensome, where in addition to mental disruption caused by unexpected prompts, participants are physically encumbered with these paper diaries, which must be on-hand for the duration of a study. Indeed, physical burden is not limited to paper-based methods; electronically administered ESM studies often require that participants carry cumbersome, bespoke study devices. Furthermore, if an assessment schedule is too intense, or does not fit with participants' own personal schedules, study attrition (drop-out) is a risk. Participants' compliance is contingent on a number of external factors, but studies tend to show a significant decrease in both data quantity and quality over as little as seven days [35]. Attrition and non-compliance are major issues that require a careful balance of assessment frequency, length, and timings that will vary between and within groups of participants. Christensen et al. propose that the intensiveness of the sampling schedule and the length of each requested assessment are the greatest contributing factors to participant compliance. As a general guideline, they suggest that each assessment should take no longer than two minutes to maximise completion rates [8].

Due to the onerous nature of experience sampling studies, it has been hypothesised that a "self-selection bias" may occur, whereby those who volunteer to participate in such studies are more likely to be motivated, technologically adept individuals [12]. Experience sampling studies may be particularly burdensome to certain populations, especially when they are administered on smartphones [13]. Elderly populations, children, and those with low socio-economic status, are potentially problematic groups for applying ESM with. Although smartphone-based assessments are designed to lower participant burden, this depends on familiarity with such devices. For example, a study by Harrison et al. showed that the uptake and adherence to a smartphone-based self-help program was inhibited by older adults, who would only use their smartphone for emergencies. Access by those in rural areas with poor network coverage was also limited [36].

Compliance is a complex and unpredictable metric, as shown by Sokolovsky et al. in their study of adolescent smokers [37]. Highly positive and negative moods had deletorious effects on compliance, as did locations outside the participants' homes. Interestingly, a greater spacing between survey requests also resulted in worse compliance. However, other studies suggest that mood may not be a determinant factor. For example, a recent study by Bos et al. showed a high compliance rate of 84.4% in depressed patients with and without anhedonic symptoms [38]. This conflicting information suggests that there are no rules that guarantee compliance.

2.1.2.2 Reliability of reports

Although experience sampling is intended to minimise the time between an event of interest and its subsequent reporting, recall bias can still be a problem. When paper methods are used, it is common for participants to complete a number of missed diary entries at the same time - a phenomenon known as “backfilling”. This issue was potently highlighted in a study by Stone et al. where participants were given paper diaries to complete, which were secretly instrumented to log when they were opened, and for how long. The results were a serious concern to the validity of previous diary studies, in that while 90% of entries were completed, only 11% were completed within the allotted 30 minute time slot. Indeed, 45% of participants “forward-filled” entries in order to alleviate themselves of future burden [39]. Periods of backfilling time were prevalent immediately before weekly updates with the study researchers.

In addition to issues of honesty, there is evidence that regularly prompting participants to answer questions on particular experiences of their life can invoke *reactivity*, where the perception of these experiences are altered by the act of their assessment. In some fields of research, this has been termed the “Hawthorne Effect” and is a significant detriment to the validity of participant research [40]. For example, if participants are frequently asked to monitor their food cravings, the constant reminder of food could make these cravings more salient, thus affecting further reports. This *negative* reactivity occurs when a participant reacts to continuous prompts in potentially harmful ways [41]. Conversely, *positive* reactivity occurs when a participant engages in healthy behaviour in response to prompts. A critical review of alcohol treatment assessment reactivity was undertaken by Clifford and Davis [42], demonstrating moderate to strong effects of repeated assessment on the decreased consumption of alcohol. Clearly this can be an advantage in treatment, in that the simple act of self-reporting a behaviour can invoke positive change.

Although participants may be out of control of their reactivity to continuous assessment, they may also deliberately change their responses in order to positively present themselves. This is a type of reactivity known as *social desirability bias*, which in addition to threatening the validity of research data, can result in significantly worse health outcomes in practical applications, for those who falsely report positive results [43, 44]. Such a bias may also be contingent on the particular area of study. For example, the study of aggressive behaviours or excessive overeating may cause participants to feel self-conscious about reporting such behaviours [45].

2.1.2.3 Practical implementation

Of particular relevance to this thesis are the practical issues associated with implementing an ESM study. Electronic administration of ESM has become a commonly accepted standard, but can place additional burden on both researchers and participants. From a hardware perspective, acquiring bespoke study devices is costly for researchers, and burdensome for participants, who are inconvenienced by carrying an additional device. Cost and burden are also inherent in the implementation of software on chosen devices. The development of ESM software can incur financial cost, whether through hiring a professional programmer, or relying on the services of a commercial company. While free software and services are currently available, many are restricted to obsolete devices, such as the *Experience Sampling Program* for PDAs [46]. Others are unsupported research projects, or have functionality limitations that impact participants' experience and thus reduce compliance.

In summary, these challenges serve to show that experience sampling is not a straightforward methodology. Nevertheless, its advantages are such that it continues to be frequently employed in psychology and clinical research. There are no rigorous guidelines to the schedule of assessments, and studies take a variety of approaches, as described in the following section.

2.1.3 Sampling strategies

The schedule of assessments is a critical factor in experience sampling studies. The strategy must be chosen to match the research question, and often combinations of strategies are utilised to maximise the chance of collecting relevant data. The sampling frequency and duration of a study must be balanced between capturing rich, informative data and minimising participant burden. The frequency of assessments may depend on the number and complexity of questions to be answered at each assessment, the number of days over which a study takes place, or the compensation given to participants for their compliance. Generic sampling strategies that do not rely on particular technology are defined by Wheeler and Reis as interval-contingent, signal-contingent, and event-contingent, described as follows [47].

2.1.3.1 Interval-contingent

With interval-contingent sampling, participants respond to prompts at specific times of the day, or at predictable, regular intervals. For example, participants may complete an assessment every morning upon waking, or receive prompts spaced two hours apart.

This sampling style is advantageous in that participants are aware of when assessments will be prompted, so are more likely to be prepared to engage in an assessment at these times. However, a disadvantage of this is that participants, in their anticipation of prompts, may pre-emptively interrupt their own streams of consciousness, and consequently lose the thoughts and feelings that assessments are intended to capture. As such, anticipation of a prompt can also cause reactivity. Further, if these intervals are spaced too far apart, they may cause recall bias [47].

2.1.3.2 Signal-contingent

With signal-contingent sampling, participants respond to prompts that come at unpredictable times. Such studies generally have participants receive a set number of prompts that are randomly distributed in their waking hours. Signal-contingent sampling is therefore more appropriate for studying phenomena that could be susceptible to recall bias, and to states that may change due to conscious emotion regulation [8].

One potential disadvantage of this schedule is that prompts can be clustered within a short space of time, increasing participant burden. To minimise this burden, prompts are often scheduled to occur once within given time windows, with a constraint on the minimum time between two prompts. Even then, random prompts are burdensome when they occur at a time that is inconvenient for a participant, such as when driving or showering.

2.1.3.3 Event-contingent

Finally, temporal sampling schedules are not always appropriate for monitoring particular events or behaviours. For example, assessing participants' experience during or directly after smoking could require an obtrusive number of prompts in order to catch the event with reasonable accuracy. Instead, event-contingent sampling has participants self-initiate assessments on the occurrence of these particular events. This is advantageous in that experiences are captured with no time delay, resulting in minimal recall bias. Further, as participants choose the times at which they complete surveys, it is the least burdensome of the sampling strategies.

One problem with event-contingent sampling is that participants may be inclined to under-report events. Although attempts are being made to capture events of interest implicitly, such studies may require participants to wear obtrusive hardware such as a wrist-worn accelerometer, or a bio-harness to monitor heart rate and blood pressure [48]. (Initiating assessments based on automatically sensed context is known as *context-contingent sampling*, and will be discussed in Chapter 3.) Event-contingent sampling is also inappropriate for capturing the antecedents and consequences of behaviours of interest, or for mapping out participants' history.

These three sampling schedules remain central to present-day experience sampling studies, whereas the technology upon which they are deployed has evolved significantly, as explained in the following section.

2.1.4 Evolution of implementation

The implementation of ESM studies has progressively taken various forms in response to technological advancements. However, despite these advancements, often the complexity of

Table 2.2: Media for ESM data collection, their advantages and disadvantages

Medium	Advantages	Disadvantages
Paper Diaries	Cheap Flexible Familiar to all participants No programming required	Cumbersome Can't monitor compliance Participants can backfill Requires transcription
PDA	Cheap Less bulky than paper diary Easily distributed Data in digital form	Cumbersome Can't monitor compliance Requires programming
SMS with online survey	Can use participant's device Real-time data transfer No programming required	Requires Internet connection Requires mobile signal Potentially expensive if multiple SMS messages are sent Tedious if SMSs must be sent manually
Smartphone	Can use participant's device Capture implicit context Real-time data transfer Sophisticated sampling strategies	Potentially expensive if participants must be loaned devices Requires programming

integrating sophisticated technology into an ESM study is beyond the programming capabilities and financial resources of researchers. Due to these constraints, they may be required to default to more simplistic or cheaper methods. Media upon which ESM studies have been employed are detailed as follows, and summarised in Table 2.2.

2.1.4.1 Pen and paper

Historically, ESM evolved as a manifestation of the diary study, in which participants were requested to complete paper diaries at regular intervals, with examples dating from the early 20th century. ESM was conceived as a way of prompting diary completion in situ, by providing participants with an electronic device that signalled them to do so. While traditional diary methods theoretically introduced the ability to monitor individual, dynamic processes, and to collect ecologically valid data, their validity is still subject to recall bias. For example, completing a diary entry at the end of a day still requires a participant to recall the entire day's events, which could itself reduce the validity of data. Further, infrequent diary entries preclude the detection of short-term, rapid fluctuations throughout the day as they unfold.

Prior to the beginning of this century, nearly all ESM studies were conducted with paper diaries. At previously set times, or on occurrence of particular events, participants would fill out paper forms and return them to researchers at the end of the assessment period. As previously discussed, backfilling of these diaries is a particular problem, with no direct means of determining exactly when the participant initiated and completed the diary entry.

Nevertheless, in their simplicity, paper diaries still have potential advantages over more advanced implementations; paper is cheap, easily replicable, and intuitive for all literate participants. From the researcher's perspective, setting up a study does not require any additional technology or knowledge of programming. However, participants must still incur the burden of carrying around these paper forms, and possibly an additional signalling device. Researchers are also burdened by this medium of data collection, as further analysis of collected data requires manual transcription of paper entries into a digital form.

2.1.4.2 Bespoke devices

Electronic ESM was reported as early as 1985, where portable "microcomputers", carried on a shoulder strap, were pre-installed with experience sampling software, distributed to participants, and returned at the end of the study for data download and analysis [49]. Intended for proof-of-concept studies, such devices were more physically burdensome to participants than paper diaries, although the central data storage and ability to timestamp entries for determining compliance was appreciated by researchers in these early studies.

As technology evolved, bespoke handheld devices such as PDAs became the state-of-the-art in ESM study devices. Much like their paper diary counterparts, these devices were loaned to participants and collected at the end of the study for data download. At the present time, these devices are cheap and as such pose significantly less financial cost to researchers. However, they may still be cumbersome to carry, particularly given that participants are likely to be carrying a separate mobile device simultaneously.

Many of the free tools allowing researchers to create ESM study applications are intended for these bespoke devices. For example, the *Experience Sampling Program* is designed to be run on PDA devices [46]. Other commercial tools such as *Satellite Forms*[1] and *Pendragon Forms*[2] are also used in PDA-administered studies. (Further tools are discussed in detail in Section 4.2.)

[1] http://www.satelliteforms.net/ Accessed 17/03/18
[2] https://pendragonforms.com/ Accessed 17/03/18

Although PDAs are still prevalent in recent ESM research, their use is declining in favour of more sophisticated smartphone applications. For example, the *PsyMate* platform initially provided software that would run on bespoke devices. However, this has now been replaced by a smartphone app for Android and iOS devices[3].

2.1.4.3 Phone/SMS services

Other common techniques that do not require a bespoke device or app are those that utilise the call and messaging functionality of participants' mobile devices. In such studies, participants are either called and asked to respond to survey questions vocally [50, 51], or sent assessments in an SMS message. In the former case, phone calls may be initiated directly by a researcher, or automatically through a pre-programmed schedule combined with an Interactive Voice Response (IVR) program. Using an SMS protocol, participants may receive the survey questions directly, to which they respond with another text message containing their answers [52, 53], or alternatively, the message provides a link to an online survey [54, 55].

If, in the case of the phoning protocol, participants talk directly to a researcher, this could be particularly sensitive to social desirability bias. Further, vocally responding in a public place is less private than completing a written diary entry. Another disadvantage to both phone and SMS protocols is that they rely on participants having sufficient network coverage, which may not be the case in rural areas. Finally, if participants are required to respond to surveys via SMS, this incurs financial burden on these participants, which they must be compensated for.

2.1.4.4 Smartphones

Smartphones are increasingly popular across all age groups and cultures, making them a viable option for deploying experience sampling applications to a wide audience. Indeed, a recent survey suggests that smartphones are beginning to penetrate the senior population, with over 70% of surveyed users aged 55-75 reportedly using a smartphone[4]. By developing apps that can run on participants' own devices, this alleviates the burden of carrying a bespoke device, or paper diaries. If native apps are used for assessment prompts and completion, this also eliminates the problem of network coverage, and financial cost of sending SMS messages to researchers.

Smartphones are powerful computers, with functionality extending far beyond the basic calling and messaging features of their predecessors. They are thus ideal platforms for providing and collecting contextual information from their owners [6]. Further advantages of smartphones are discussed in the following chapter, and in particular their contribution to state-of-the-art

[3]http://psymate.be/psymate–en-.html Accessed 17/03/18

[4]Source: https://www.deloitte.co.uk/mobileuk/#hi-ho-silver-swiper. Accessed 17/03/18

ESM studies. As a short summary, sensor capabilities, web connectivity and widespread mass adoption are some of the strongest advantages. Smartphones contain a wealth of sensors and data streams that can be probed to provide additional contextual information, independent of external hardware or user input. Additionally, smartphone web connectivity allows data to be instantly accessible by researchers as a participant records it, such that compliance can be monitored and data analysed in real-time.

2.2 Recent ESM Studies Review

In order to determine the current state of ESM in research, a literature review was conducted, with a similar structure to that of van Berkel et al. in a recent survey of ESM on mobile devices [56]. Their systematic review used digital libraries focused on computer science literature, providing an excellent overview of ESM in this domain. Instead, the following review is intended to assess the use of ESM in general research, where the focus is on participants as opposed to the technology used to assess them. This review is not systematic - to select a tractable number of publications with the broad scope of available literature, significant constraints had to be applied. For example, the search term "*'experience sampling' OR 'ecological momentary'*" returns 17,700 hits on Google Scholar, with 7,320 since 2017 alone.

The final search was modified to only include publications with *experience sampling* or *ecological momentary* in the title, limiting the returned results to 318 since 2017. With this constraint, the following review only covers a small fraction of relevant publications. For a systematic review, in addition to surveying all 7,320 results returned in Google Scholar, other digital libraries would have had to be used, as well as additional keywords. For example, *ambulatory assessment*, a phrase often applied to studies recording physiological measurements, can also represent studies that repeatedly collect self-reported data [26]. Nevertheless, employing this filter means that a broad range of literature is covered, including publications from psychology, medicine, computer science, and business journals, while keeping reviewed publications to a reasonable number. Further, this strategy ensured that experience sampling was a focal point of the publication. Studies were removed if they were meta-analyses of ESM, described ESM as a potential research method in a particular domain, proposed a protocol for a randomised control trial, or the full text of the publication was unavailable for access. After this filtering process, a total of 121 publications remained.

Following this, the same protocol was repeated, using Google Scholar to find publications from 2006-2008, to cover the period around which Intille published a visionary paper of experience sampling in 10 years [57]. This provided insights into how experience sampling has evolved in this 10 year period. From the search process, a total of 177 hits were returned that, after the same filtering process, resulted in 43 publications being included, for a total of 164 publications. (However, there are 165 studies due to one publication containing two separate ESM studies.) A table of these publications with their parameters and references is included in Appendix A. The broad research area of each of these publications is almost exclusively psychology or medicine, with all studies designed to capture participants' emotional or physical states across a set time period. In order to represent the spread of publications more specifically, high-level categorisation was employed, based on publications' titles and abstracts. For example, 29 studies focus on participants with addiction to drugs, alcohol or smoking. 18 studies focus on participants engaging in weight management, where diet or exercise habits are tracked. 10 studies focus on participants tracking the symptoms of physical impairments and diseases, including HIV and cancer. 48 studies involved participants with varying degrees of specific mental disorders, including depression (11), eating disorders (10), schizophrenia (9), psychosis (7), borderline personality disorder (4), post-traumatic stress disorder (4) and ADHD/autism spectrum disorders (3). In addition to such studies published in psychology and medicine journals, six were published in the domain of education (with four related to students' learning and engagement, and two related to teachers' stress and emotional experiences). Finally, of particular relevance to this thesis, three publications address HCI issues towards increasing compliance in ESM, including visualisation [58], gamification [59] and use of wearables [60].

The systematic review by van Berkel et al. of studies in computer science (hereby referred to as "CS studies") collected various parameters, on which the studies included in this review are compared, including: number of participants, duration, compliance, compensation, trigger type, use of sensors, and device ownership. In addition, the type of device used (ranging from smartphones to paper diaries) is included. Finally, the total number of assessments prompted over the course of a study is also included to assess the level of impact this had on compliance.

2.2.1 Study parameters

The following section summarises the reviewed studies with respect to the parameters listed. Although significant effort has been made to ensure accuracy of results, only the author reviewed these studies, so it was not possible to resolve ambiguities and determine cross-researcher agreement. Further, although it was expected that correlations between participant compliance and other study parameters could be investigated, the lack of information given in many publications means that any significance tests conducted lack statistical power. For example, 24 studies (14.5%) do not report on participants' compliance, whereas 51 (30.9%) do not report on the compensation given. In two studies, the prompting schedule is also absent.

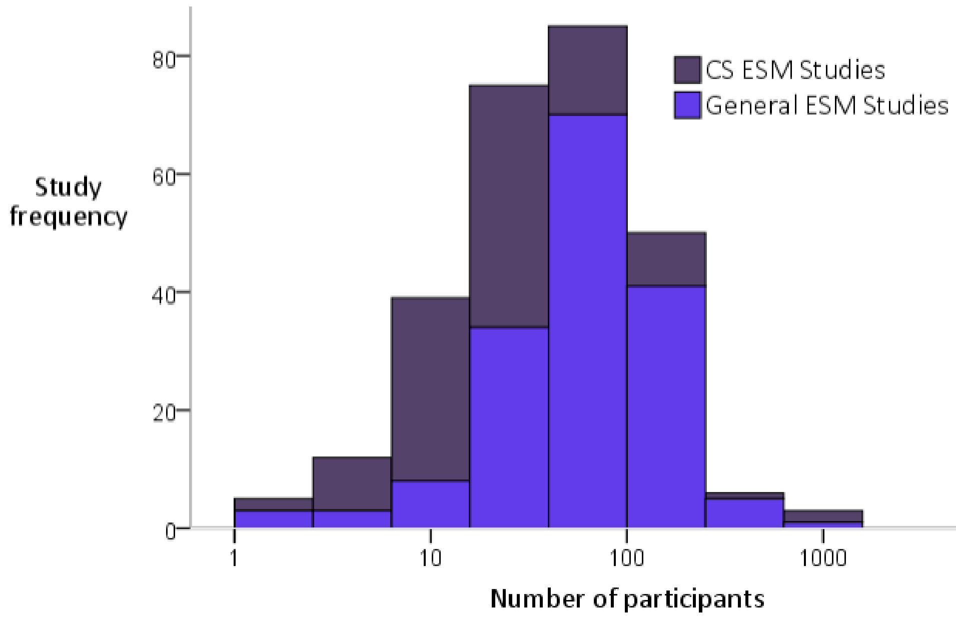


Figure 2.1: Participant numbers of CS studies and general ESM studies

2.2.1.1 Number of participants

In the 165 reviewed studies, the mean number of participants is 89, with a median of 59. This is considerably higher than the mean of 53 and median of 16 reported in CS studies. Journals in psychology and medical research generally require study samples to be large to ensure a high statistical power, unlike human-computer interaction research where ESM studies are more likely to be proof-of-concept ideas or small-scale evaluations. The implication of this need for larger sample sizes is that considerably more effort is required for recruitment, data retrieval, analysis and debriefing, especially given that a minority of recent studies still employ PDAs or paper, requiring participants to return to a lab for data transfer. If participants' own devices could be used to download an app, then it may be possible to recruit, assess and debrief large numbers of participants remotely. The spread of sample size for both sets of studies is shown in Figure 2.1.

2.2.1.2 Study duration

Interestingly, the duration of general ESM studies is considerably lower than that in the computer science literature. CS studies have a mean duration of 32 days and a median of 14 days, whereas studies in this review have a mean duration of 15 days, and a median of 7 days. Indeed, the majority of studies take place over the course of one week. This is to be expected, given that guidelines on conducting ESM procedures recommend that studies should be kept short to increase compliance [8]. Despite this recommendation, there does not appear to be a correlation between the number of study days and compliance of participants in the studies in general

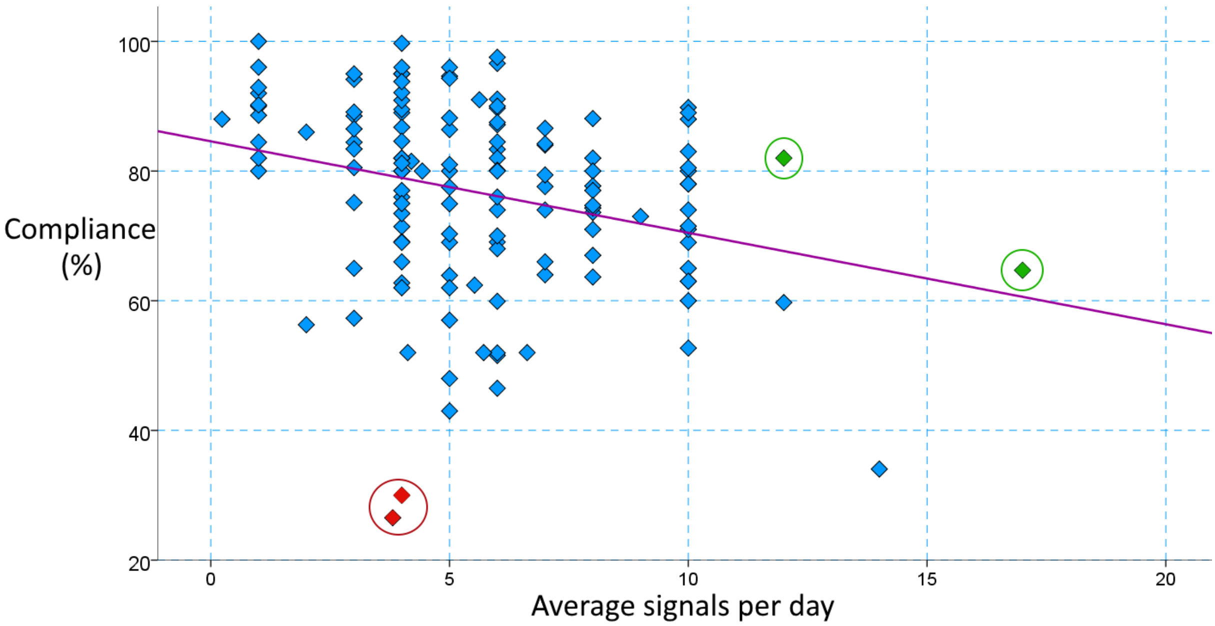


Figure 2.2: Study compliance tends to decline with sampling intensity

literature, or those in the CS studies. A Pearson product-moment correlation coefficient showed no significant correlation between the two variables $(r = 0.0062, n = 141, p = 0.467)$. However, compliance is dependent on many other factors including the intensity of daily sampling, and whether compensation is awarded.

2.2.1.3 Number of prompts

While the duration of a study does not appear to significantly influence the compliance of participants, the intensity of studies (determined by the average number of prompts per day) is a stronger predictor of compliance, as discussed in the next paragraph. Intense sampling schedules are often required to capture variables such as activity, where implicit measures are not available, as in [61]. In other studies with intense sampling, external devices were worn (such as actigraphy accelerometers, ECG monitors or light sensors). While reasons for intense sampling in these studies were not stated, it may be that the required number of samples had to be obtained in fewer days to minimise the burden of long-term wear of an external device. For example, one study utilising a body harness that took physiological readings was conducted for only 36 hours, but prompted 15 self-reports [62].

2.2.1.4 Compliance

Participant compliance varies widely, with no statistically significant correlation to any single parameter such as study duration, device type, device ownership, or total number of assessments.

However, a significant correlation was found between compliance and the average number of signals a participant received per day. A Pearson product-moment correlation coefficient was computed to assess the relationship, showing a significant negative correlation between the two variables $(r = -0.280, n = 137, p = 0.001)$. Figure 2.2 illustrates this negative relationship.

However, as illustrated there are other potential factors that contribute to a high or low response rate. Consider the circled studies in the lower-left of Figure 2.2, which have low average compliance despite a small number of signals. In one of these studies [59], participants responded to a signal in which they were asked to engage in a task of assigning a relevant keyword to a location, followed by rating keywords assigned by others. This task is likely to be longer and more cognitively demanding than simple questionnaires. The other study was conducted with post-hospitalisation participants with psychotic-spectrum disorder, who are also often poor at adhering to their medication [63], and thus were not expected to be compliant with a study for which there was no penalty in failing to do so.

Conversely, two studies had particularly high compliance with respect to their sampling intensity, circled in the upper-right of Figure 2.2. In one of these studies, participants were police officers, the study took place over a single day, and the reward was a raffle for a large monetary prize [64]. In the other, participants were adolescent females who were prompted at weekends, hence less likely to have school or work commitments. They were also paid per survey they completed [65].

Furthermore, even with a significant correlation between compliance and sampling intensity, the variety of different app interfaces, participant demographics, assessment length, time limits, as well as the frequent inadequate reporting of such information, make it difficult to accurately predict compliance.

2.2.1.5 Compensation

Compensation is often unreported in both CS studies, and the studies of this review. Some report that participants were *"compensated for their time"* or *"received payment"*, but with details omitted. Many studies do not report whether participants received compensation at all, such that poor compliance may be a consequence of no financial motivation. Compensation mechanisms in the reviewed studies include unconditional fixed rewards, levelled rewards based on compliance, a raffle system, or completely voluntary participation. With the small sample size and variation in reward levels, it is impossible to determine whether one mechanism is more effective than another, and for whom.

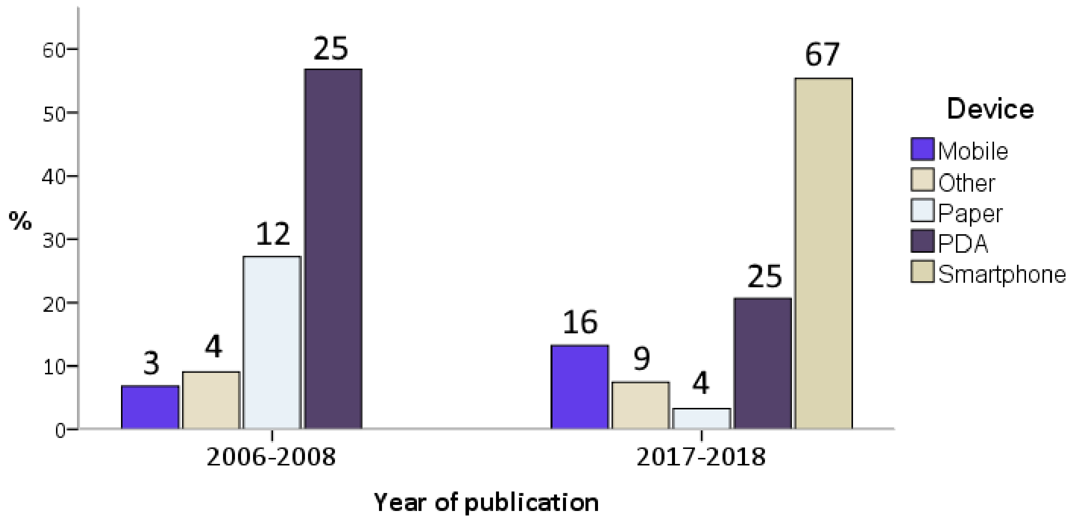


Figure 2.3: Range of devices used in the 2006-08 sample and the 2017-18 sample

2.2.1.6 Sampling schedule

Schedules are either described as signal, interval or event-contingent [47]. Of the 165 total studies from both time periods, 86 (52.1%) used a purely signal-contingent schedule. The next most popular sampling schedule was interval-contingent, with 33 (20.0%) studies, followed by a signal/event-contingent combination, with 18 (10.9%) studies. Compliance with event-contingent schedules is difficult to determine, given that participants initiate assessments in response to their own subjective interpretation of events. Exceptions are studies that compare explicit self-report with an implicit measure of events occurring. For example, one study compared self-reported smoking behaviour with that collected by a Bluetooth-enabled e-cigarette [52]. Only one of the general ESM studies, incidentally reported in a computer science journal, used an app that triggered assessments based on location [59], and no study had assessments triggered on any other form of implicit context.

2.2.1.7 Sensor usage

Unlike the CS studies where 63.6% collected some form of sensor data, only 14 of the 121 studies sampled from 2017-18 (11.6%) reported doing so. The majority of these studies used either an external device (such as a wrist-worn accelerometer) or had participants install a separate app to track location or activity. Thus, despite ambitious predictions of 10 years previous [57], advocacy of smartphones for psychology research [6] and the wealth of available sensor data [7], only a very small minority of reviewed studies exploit smartphones' sensor capabilities.

2.2.1.8 Device type

The proportions of devices used across both early and recently sampled studies are illustrated in Figure 2.3. In the 2017-18 sample, as expected, the majority of ESM studies were conducted with a smartphone (68, or 56.2%). However, PDA devices were also used in a significant proportion (25, or 20.7%), and four studies even reported the use of paper diaries. For smartphone studies, most reported use of a native app (51), while the remainder had links to online surveys sent to participants via SMS (17).

Of the 18 studies in the *Mobile* category, where participants used mobile devices that were not explicitly identified as smartphones, the most common protocol involved SMS or app-based solutions (11, or 61.1%) . A minority instead involved researchers calling participants at random times of the day (2), or using Interactive Voice Response (IVR) software (5).

The *Other* category includes devices that were neither mobile phones nor PDAs, such as iPod or iPad devices, smart watches, or studies where participants were simply emailed ESM assessments and asked to reply on any device of their choosing.

2.2.1.9 Device ownership

Figure 2.4 illustrates a comparison between the ownership of study devices in the 2006-08 and 2017-18 samples. This chart highlights that, while the use of personal devices has greatly increased in 10 years, general ESM studies still largely employ bespoke devices for the duration of a study.

2.2.2 Limitations

As previously discussed, the review process was not systematic. Studies containing the terms “experience sampling” or “ecological momentary” in the title are not necessarily a representative sample of all ESM studies in their particular years of publication. Indeed, it is possible that studies containing these terms in the title are more likely to be feasibility studies of the method, given that many of the reviewed publications were presented as such.

Despite this concern, the studies were undertaken with a wide variety of participant samples and varying degrees of complexity. Compliance rates, study duration, and sample size are consistent with systematic reviews of ESM studies conducted in specific domains [66, 67].

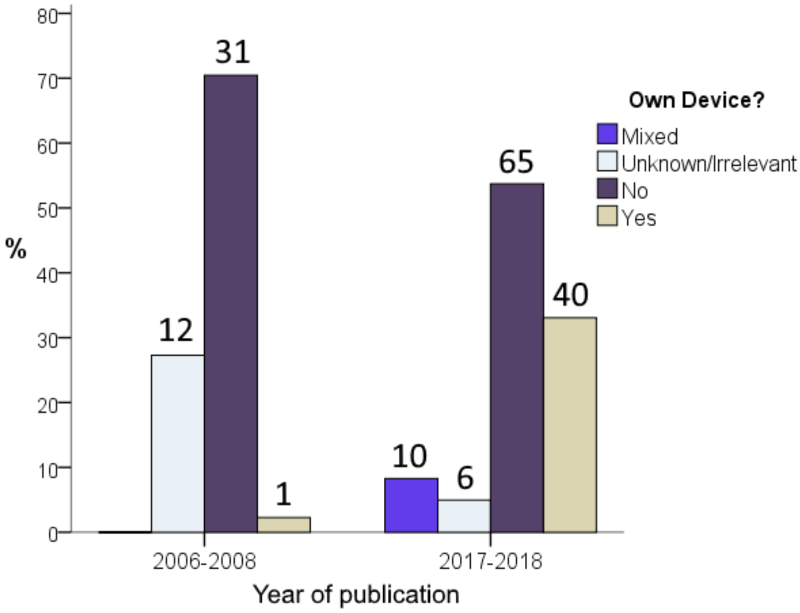


Figure 2.4: % device ownership in general ESM studies. Absolute numbers appear above bars

2.2.3 Discussion

Although there are many similarities between the review by van Berkel et al. in [56] and this review, some notable differences were identified, particularly regarding less device ownership in general ESM studies. It is probable that this is a consequence of the recruited samples; many CS studies recruit university students via convenience sampling, whereas studies in psychology and medicine recruit a variety of participants with different ages, education levels, and disabilities.

With exception of smartphone usage, there are very few differences between the reviewed studies published 10 years apart. Instead, researchers appear to be running the same type of simplistic studies, but on devices that are capable of more. One reason for this could be the difficulty or resources involved in programming more complex apps. Without available tools to develop such apps, researchers may be forced to rely on outdated technology, or on simple apps that take minimal time and cost for professional developers to implement. It was notable that six of the studies in the 2017-18 sample reported using the aforementioned *Experience Sampling Program* software to program PDAs for participants.

Indeed, it is not to be discounted that researchers and clinicians simply may not wish to do anything more complex. An ESM study design is dependent on the research question it aims to answer, for which a replication of a previously successful study protocol may be sufficient. However, the following chapter provides evidence that more complex functionality would be particularly useful for researchers, clinicians, and their participants. Regarding replication, it was notable in both this review and that of van Berkel et al. that there is often an inadequate description of methods to ensure that researchers can replicate studies. When compliance rates, study compensation, and any further explanation of app design are absent from publications, this raises questions of ambiguity when other researchers fail to replicate significant findings.

2.3 Implications

The findings of this chapter's literature overview and survey of previous ESM studies motivate the following implications for the design of Jeeves.

Foremost, ESM is inherently burdensome for participants, requiring them to engage in assessments through the course of their everyday lives. Burden is manifested both *mentally*, if participants are prompted at times where their tolerance for interruption is low, and *physically*, if they are required to carry around an additional device for a study's duration. The survey showed that compliance is negatively correlated with sampling intensity, and that many recent studies are still being conducted with bespoke devices.

Second, ESM can be burdensome for researchers conducting studies using the method. The survey showed that in general, sample sizes are high to reduce the risk of bias. Consequently, monitoring compliance and managing data analysis becomes non-trivial, particularly if bespoke devices have to be acquired.

Finally, the variation in participants over the different studies (including children, mentally ill patients, and the elderly) coupled with the unpredictable variation in compliance statistics, demonstrates that there are no static rules that guarantee a study's success. However, the ability to tailor to the needs of individual participants over the course of a study could address this issue.

2.4 Chapter Summary

This chapter has described the Experience Sampling Method in more detail, as a longitudinal research method in psychology, as well as a possible means of supporting treatment and intervention strategies in clinical practice. Smartphone-based ESM implementations afford new opportunities for both research and practice, with an emphasis on providing real-time, contextually relevant information to both researchers and participants.

Despite the promises of smartphones for facilitating remote longitudinal assessment, the actualisation of these promises has posed significant challenges to researchers wishing to do so, as summarised in the implications described previously. In particular, barriers to implementation have hindered potential new insights in psychology and medicine - a problem that underlies the primary research question of this thesis. The following chapter expands on recent progress in smartphone-based ESM, which this thesis proposes could address the issues represented by these implications.

CHAPTER THREE

SMARTPHONE DATA COLLECTION

The benefits of ESM to research were discussed in the previous chapter, ending with a timeline of technological developments that have facilitated the method's use. The first section of this chapter expands on this further, describing the recent developments in ESM afforded by smartphone technology. Further, example applications of smartphones in psychology research and clinical practice are described, framing the context.

This is logically followed by a description of smartphone ESM, with particular emphasis on the needs as well as desirable features for researchers and participants/patients to ensure consistent, long-term usage. These features are synthesised into a model of smartphone ESM interaction, which provides a frame of reference for the current state-of-the-art in ESM creation tools.

3.1 Smartphone ESM Developments

Beyond standard ESM with simple time-based sampling schedules, the widespread adoption of powerful smartphones has afforded benefits previously unavailable in applications of the method. For example, smartphone ESM is applicable to **participatory sensing** projects, in which study participants use their own devices to provide data towards a large-scale research goal. Further, smartphones are inherently personal devices, storing a wealth of information that could be used to tailor applications to their users. In the domain of ESM, self-report assessments provide rich insights into participants' physical and mental well-being, in order to inform **ecological momentary interventions**. Finally, sensors built into the smartphone itself such as GPS sensors, or worn externally, such as wristband activity trackers, can automatically collect rich contextual information on their users. In ESM apps, this context can be used to prompt assessments or interventions, conceptualised as **context-sensitive ESM**. These three innovations will be discussed in turn.

3.1.1 Participatory sensing

The prevalence and connectivity of sensor-rich smartphones extends their capabilities far beyond the individual level, into something of a global sensor network, through which sensing applications can collect large quantities of data. In participatory sensing studies, participants are often individual data sources within a larger project that aims to answer research questions on a community, national, or even global scale. In some cases, a participant simply has to download and run an app, which autonomously collects relevant data. The simplicity of contribution to the altruistic goals of these large-scale studies, has participants acting as *"citizen scientists"*, often supplying data for a collective good rather than their own personal benefit [68].

As opposed to traditional ESM studies in psychology, the focus of a participatory sensing study is on an individual's environment, rather than the individual themselves. This is the primary difference between the two types of study, such that participatory sensing studies are often implemented for studying environmental or sociological issues. The development of sophisticated sensor capabilities enable participants to have different levels of involvement in the data collection process, defined in terms of two types of "crowdsensing" as follows:

- **Participatory crowdsensing** has participants decide the appropriate conditions under which to engage in a study. Although sensor data may be acquired, it is primarily obtained through self-report surveys, and thus requires participants' active, conscious involvement in a crowdsensing application.

- **Opportunistic crowdsensing** applications run in the background, choosing when to passively collect data, such that participants can be unconscious of any sensing that takes place. As an example, complex classifications of participants' mobile sensor data was used to predict the expected arrival of buses [69].

In general, participatory sensing studies are more flexible than ESM studies in order to minimise burden on participants, given that they do not necessarily receive financial compensation for their participation. For example, time-contingent sampling is seldom employed, instead relying on participants to engage in manual event-contingent sampling. Further, additional incentives must be provided to the participant in lieu of compensation. A recent review of non-monetary incentive mechanics in health-based participatory sensing apps was conducted by Anawar et al [70], in which self-monitoring features were found to be the most reliable in maintaining

Table 3.1: Participatory sensing incentives and their applicability to ESM studies

Incentive	Definition	Applicability to ESM
Monetary	Participants rewarded with money	Current practice. Not sustainable for large sample sizes
Collective	Participants motivated by attainment of a 'common good'	Not useful. Purpose of ESM study unlikely to be motivated by community benefit
Social (interaction)	Social networks leveraged, allowing participants to share and interact	Not useful. Privacy issues prevent sharing and interaction with other participants
Social (self-interest)	Participants receive collective info through contributing: 'quid pro quo'	As above
Monitoring	Participants empowered to monitor and manage their own data	Potentially useful, participants could be incentivised by monitoring study progress
Autonomy	Participants in full control of what they send, where and when	Potentially useful, if scope for flexibility in data collection process is feasible
Feedback	Participants can send and receive feedback on their progress	Potentially useful, ensuring feedback both ways is kept private and anonymous

engagement. This result is consistent with behaviour theories that support individuals' perceived autonomy to make choices and exert influence as a motivational factor [71, 72].

In their application to understanding social and environmental conditions, participatory sensing apps are often employed in fields of research such as social science, where researchers are less likely to have significant programming experience. As such, tools to allow creation of participatory sensing apps are often a necessary prerequisite for such research.

3.1.1.1 Authoring Tools

A number of existing tools allow researchers to orchestrate "participatory sensing campaigns". EpiCollect [73] and Open Data Kit [74], for example, are tools for designing questionnaire forms that can be deployed onto participating users' devices. They allow a variety of question types, and multimedia data inputs (such as audio, video and photo) with optional branching and skip logic. Researchers and participants can view their collected data in the form of graphs and charts on a project website. However, such applications still rely on participants' incentive to initiate the surveys themselves, limiting any form of ESM applications to event-contingent sampling schedules. Furthermore, the complexity and diversity of sensor classifications would make it impossible to allow non-programmers to orchestrate *opportunistic* crowdsensing apps.

Nevertheless, participatory sensing is still of relevance to the work of this thesis. Given the burdensome nature of experience sampling studies, and the financial costs to the researcher in scaling up to larger populations, participatory sensing research describes additional means of minimising participant burden and improving experience that could tentatively be incorporated into appropriate ESM studies. In particular, self-monitoring incentive mechanisms appear most appropriate for application to experience sampling: allowing participants to monitor their collected data, ask questions and receive feedback, and tailor their data collection preferences, could motivate compliance. Table 3.1 thus summarises the various types of incentive that have been applied in participatory sensing studies, and their applicability to ESM studies.

3.1.2 Ecological Momentary Interventions

Chapter 2 briefly discussed the benefits of ESM to treatment monitoring, and the previous section described the benefits of smartphones for both providing reassurance to fragile patients, but also for motivating self-management in participants with chronic conditions. Such applications of ESM are described as "*Ecological Momentary Interventions*" (EMIs) - an extension of the standard ESM approach that provides mobile users with in-the-moment feedback, advice or treatment as they go about their everyday lives. Thus, rather than data collection tools, EMI apps are intended as behaviour change tools, providing health benefits to their users. Heron and Smyth provide a thorough review of mobile EMIs, including their advantages and challenges [75]. Published in 2010, this review predates the widespread proliferation of smartphones, thus many new EMIs have been implemented that utilise the benefits of such technology.

The potential of smartphone technology to influence positive behaviour has prompted a growing number of novel applications in recent literature. Indeed, their prevalence has motivated publication of reviews and guidelines for future interventions in specific domains. A review by Spruijt-Metz et al. reviewed the state of smartphone-based interventions to support weight management [76], and Bort-Roig et al. conducted a systematic review of smartphone technology for influencing physical activity [77]. While smartphone interventions are generally positively received, reviews show that results are modest and tend to vary in significance, such that further research must be conducted into utilising the full capacity of smartphones. Two potentially beneficial factors are discussed below - smartphone sensing, referred to as *implicit context*, and individual tailoring, or *explicit context*.

3.1.2.1 Sensing - Implicit Context

One beneficial factor not previously available in the EMIs of Heron and Smyth's review, is that of smartphone sensing. In capturing participants' locations, activity levels, presence of others, or even limited physiological data, feedback based on this implicit context has potential to significantly increase the effectiveness of interventions. The following two examples suggest potential uses of capturing participants' context for the purposes of intervention.

Burns et al. present an early example of applying contextual sensing to intervention. Their *Mobilyze!* app captured a range of sensor data including accelerometry, nearby Bluetooth devices, GPS location, and ambient light. Machine learning algorithms were then trained to predict participants' moods from correlated explicit self-report data. These predictions, coupled with visual feedback charting mood states against sensed data, were well-received by participants, providing insights into contributing factors to their depressive states.

A more simple, yet effective application of implicit context is described by Naughton et al. [78] in a study employing the *Q Sense* smartphone intervention. The app asked participants to log locations of smoking behaviour, in order to deliver tailored, location-sensitive intervention feedback when the participant revisited and dwelt in this location. Although the study was presented as a proof-of-concept with no significant results, participants found the location-triggered messages to be motivational, and a helpful distraction.

3.1.2.2 Personalisation - Explicit Context

Despite the advantages of instant feedback implemented in EMIs, the effectiveness of these responses could be reduced somewhat if they are not tailored to the recipient. A randomised control trial by Kramer et al. involved a control group, a second group participating in ESM, and an experimental group who also participated in ESM while receiving weekly feedback based on their data [79]. By providing personalised insight into patterns of positive affect, they were able to find a clinically relevant and borderline significant reduction in depressive state in the group receiving ESM with feedback, compared to the group receiving only ESM.

As discussed in Section 3.1.1, behaviour change theories from psychology support personalisation as a strong motivator for engaging in positive behaviour. For example, *autonomy* is a primary construct in self-determination theory [71] and Motivation 3.0 [72]. These models are useful for understanding the role of different components in an individual's behaviour, and thus they provide incentive towards developing systems that conduct intervention strategies at a personal level. Given that today's smartphones are inherently personal devices, capturing life as it is experienced, they are an ideal platform on which to deploy intervention feedback. A number of reviews have highlighted the effectiveness of doing so. For example, Klasnja and Pratt reviewed mobile phone health interventions, detailing sources of potential benefit based on a number of significant results [34]. The personal attributes of participants, ranging from stable factors such as age group, to dynamic factors such as self-reported anxiety, are **explicit context**.

3.1.3 Context-sensitive ESM

Beyond its application to intervention, participants' context, captured either through sensed data (implicit context) or through results to previous surveys (explicit context) can have additional benefits to both researchers and participants.

3.1.3.1 Correlational data

The benefits of smartphone sensing for capturing location, activity, or social behaviour, go beyond EMIs. For example, passive data capture reduces participant burden by minimising the need for manual input. Collecting accurate location data with GPS, in addition to further contextual information from ESM, is also beneficial for understanding how context influences behaviours of interest, such as alcohol consumption [80]. It is also possible to infer social contexts through simply sensing the number of nearby Bluetooth-enabled mobile devices, as employed in the *EmotionSense* project [81, 82]. Thus, even without direct intervention, simply collecting implicit, correlational data could be highly informative for researchers.

3.1.3.2 Minimise interruption

Sensed implicit context also allows for more intelligent sampling strategies. For example, in addition to signal, interval and event-contingent schedules, prompts can be delivered to a participant in response to classifications of sensor data. For example, if an event of interest can be implicitly inferred with sensors, prompts could be delivered on these occasions to minimise irrelevant interruptions. Kirchner and Shiffman give a detailed report on the application of GPS location to this concept, through *"geographically-explicit ecological momentary assessment"* [83]. It is often the case that non-response from participants is caused by prompts occurring at an inconvenient time. For example, a participant may be showering, driving, in a meeting, or performing other such tasks where interacting with their device would be considered dangerous or unacceptable. Several studies have measured factors that predict a user's "interruptibility" - their state of being able to respond to prompts. This transition can be detected by smartphone sensors, and studies have suggested that instances where users are moving between tasks are opportune moments for interruption. For example, Ho and Intille prompted participants based on sensor-monitored changes in physical activity [84]. A study by Poppinga et al. found that instances where the user is holding the device, such as after sending an SMS, were appropriate times to send notifications [85].

Table 3.2: Selected rows of table in [57], describing the state of context-sensitive ESM

Property	**PDA+sensors**	**Smartphones (predicted)**	**Smartphones (reality)**
Lightweight	🟡	🟢	🟢
Compact	🟡	🟢	🟢
Inconspicuous	🔴	🟢	🟢
Easy to replicate	🔴	🟢	🟢
Easy to modify assessment strategies	🔴	○	🟡
No special expertise required to implement	🔴	○	🟡
Subject can adapt if desired	🔴	🔴	🔴
Auto-tailor prompts and questions to context	🟡	🟢	🟡

🟢=property satisfied 🟡=property satisfied to a lesser degree 🔴=property not satisfied ○=uncertainty

3.1.3.3 The vision of Context-Sensitive ESM

Context-sensitive ESM was reviewed by Intille in 2007, many years prior to the general adoption of smartphones [57]. At the time of this publication, studies utilising context-sensitivity had participants carry bespoke PDA devices with additional, externally worn hardware. His chapter proposes advantages of future mobile devices for running experience sampling studies, and in doing so, constitutes a visionary publication for the research conducted in this thesis. Summarising his predictions in a table, Intille explains:

"*...researchers interested in studying free-living subjects may have a powerful new assessment instrument at their disposal as the technology that enables widespread adoption of [context-sensitive ESM] matures*" [57, p. 332].

Selected rows of Intille's table are replicated in Table 3.2, with an additional column added to show how his predictions compare to the current state of context-sensitive ESM. Although correct in the majority of his predictions, there are some that remain unfulfilled, and the highlighted cells represent the deficiencies that the work in this thesis aims to address.

3.2 Introducing ESM to Research and Practice

The previous section discusses the advances that smartphones have enabled in the application of ESM in research and practice. In order to substantiate the proposed advantages of these developments, literature was reviewed that presented feedback on research and healthcare apps from the perspectives of researchers, clinicians, and participants.

The following literature search was conducted within computer science literature databases including the ACM, IEEE, and Scopus digital libraries, as well as the PubMed Central and APA PsycINFO databases, using combinations of terms including ‘smartphone’, ‘self-monitoring’, ‘experience sampling’, ‘ecological momentary’, ‘user-centred’, ‘participatory design’ and ‘recommendations’. Particular insight was obtained from user-centred design studies that highlighted stakeholder perceptions of smartphone apps for self-management, prompting the addition of the latter three keywords. Arguably, user-centred design is more applicable to healthcare, given that psychology research apps are seldom intended to be of long-term benefit to participants. However, implementing features perceived as useful by patients could also improve participant compliance in research studies.

The following section conceptualises the activities of researchers, participants, and ESM apps themselves, as interactions. These three entities can interact with each other, and all can interact in different ways with participants’ data. Sequential executions of these interactions demonstrate potentially beneficial use cases as described in the reviewed studies, of which examples are subsequently provided. While some of these interactions are present to differing degrees in bespoke apps, the core motivation of this thesis is that these interactions should be enabled as necessary by the researchers themselves.

3.2.1 A model of interactions

The interactions of researchers shape the course of their specific smartphone ESM study. Moreover, they constitute a model of interactions between participants, researchers, apps and data, as shown in Figure 3.1. In this model, the researcher’s interactions are represented by arrows G, H and I, and participants’ interactions by arrows D, E and F. At the centre of this model is the ESM app, which mediates the various interactions between researchers, participants and their data, and also has its own interactions with these three entities, represented by arrows A, B and C. These interactions are described in detail, with implications for smartphone ESM.

3.2.1.1 ESM App Interactions

An ESM app is more than a static diary, and can autonomously initiate interactions with the participant through prompting of assessments. However, ESM apps can also interact with a researcher through wireless connectivity, and with a participant’s data through on-board sensor classifiers. These interactions will be explained as follows.

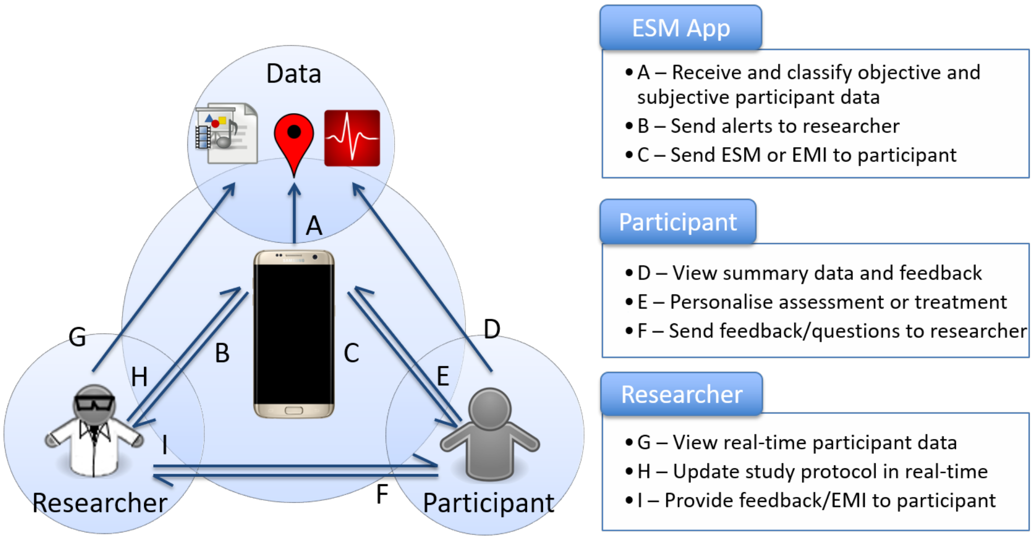


Figure 3.1: A model of interactions between researchers, participants, and ESM apps

Interaction A - Receiving and automatically classifying data

Self-reports can assess intentions, attitudes or certain explicit physical symptoms (such as discomfort or pain). However, an increasing breadth of information can be implicitly assessed from sensors built into the smartphone, or externally worn. In doing so, passively collected sensor data can minimise self-report burden, provide researchers with richer information, and enable tailored assessment and intervention strategies.

In psychology research, the UBhave project is investigating how, in combining implicitly sensed data and explicit self-report, apps can learn and classify sensor readings that relate to particular behaviours, such that interventions could be automatically triggered at appropriate times [86]. With the burgeoning interest in machine learning, the sophistication of classifications will continue to grow. Google has available Android APIs for accessing its complex, resource-efficient classifiers. Further, other high-level contexts such as the presence of others, level of sociability, quality of sleep, or susceptibility to interruption have been inferred from sensor classifiers instrumented in the computer science literature. For example, Ben-Zeev et al. showed that smartphone-sensed geospatial activity and sleep duration were significantly related to stress levels in a cohort of young people [87]. Adams et al. review smartphone sensing for monitoring physiological and behavioural biomarkers, which is a comprehensive source of further examples [7].

Interaction B - Sending alerts to researcher

From knowledge acquired through explicit and implicit data, ESM apps in healthcare could inform clinicians of in-the-moment assessment issues, or emergency situations. In many conditions, early detection of symptomatic signs can prevent fatal consequences, or prevent relapse in psychological disorders, through clinician intervention.

Given that social psychology research often involves sensitive participants with learning difficulties or psychological disorders, it is ethically important to ensure that these participants have a means to request support. For example, the potential for apps carried by dependent individuals to send alerts to carers in critical situations could assist in the independent living of these people.

In a user-centred study on app requirements, clinicians and cancer patients both supported a feature to alert clinicians when patients explicitly reported high levels of pain [88]. Clinician alerts were also implemented in a randomised control trial, where physiological readings transferred to smartphones via Bluetooth, combined with explicit symptom reports, alerted clinicians to heart failure pre-conditions [89]. The trial reported improved health outcomes, with patients expressing feelings of reassurance and self-efficacy, and clinicians confirming the utility of receiving such alerts.

Interaction C - Sending notifications to participant

As described in Chapter 2, ESM sampling schedules can be signal-contingent, interval contingent, or event-contingent [47]. In addition, sensor data classifications provide implicit context, such that **context-contingent** sampling can be employed. Assessments could also be triggered on explicit context - if participants give certain responses it could be indicative of an interesting situation, which further assessment would then acquire the details of.

In considering conditions under which to send notifications to participants, it is also important to consider the conditions under which to *defer* notifications. Participant burden continues to limit the utility of ESM, where continuous prompting in everyday life can reduce compliance when these prompts occur at inopportune moments [12]. Sending notifications to participants to complete surveys is a core component of ESM apps in research. In smartphone-based implementations, three important types of notification have been identified, namely *assessments*, *interventions*, and *reminders*, examples of which are shown in Table 3.3. Triggers for these prompts are further divided into implicit context and explicit context, further examples of which were described in Section 3.1.2.

Table 3.3: Types of notifications, their contextual triggers, and examples

Communication	Trigger	Example
Assessment	Implicit	User is detected to have engaged in high levels of activity - send assessment to ask about exercise behaviour
	Explicit	User self-reports abnormally high glucose reading - send follow-up assessment to gather more information
Intervention	Implicit	User is near location of regular smoking behaviour - send user motivational message
	Explicit	User self-reports high levels of distress in assessment - send user support information
Reminder	Implicit	User has arrived home - send a reminder to take medication
	Explicit	User has not completed required number of assessments - send a reminder to complete those missed

Assessment

Explicit context refers not just to the answers given in completed assessments, but also those that were missed. Based on participants' non-responses, further assessments can be sent. Particularly in psychology, researchers rely on continued compliance with data collection, such that the ability to send reminder prompts when self-reports have not been completed is a useful feature, and commonly applied in bespoke ESM apps. In discussing an app for obesity management, clinicians recognised the utility of such a function:

"Maybe the app could provide reinforcement messages that popup, like, 'you tracked for 5 days in a row!', 'We haven't heard from you in a while, let's check in?"' [90, p. 812]

Given the need for consistent completion of assessments, inferring the *interruptibility* of ESM app users through passive data collection is thus an ongoing research effort. In the first study assessing mobile interruptibility, an externally worn triaxial accelerometer was used to prompt self-reports following a sudden drop or increase in movement, revealing increased receptivity in these states [84]. In the clinical domain, Sarker et al. collected data and inferred higher level information from a variety of sensors, correlating it with response rate to ESM surveys, to derive a prediction model of interruptibility with 77.9% accuracy, showing how implicit data can also be used to maximise ESM engagement [91].

Interventions

Interventions can be dynamically delivered at locations of interest or on physiological measures. For example, the *Q Sense* app deployed tailored interventions when participants dwelt in identified smoking locations, with post-study feedback exhibiting positive response towards location triggers, and tailored messages [78].

One study combined results from implicit context classification, and explicit self-report to deliver an EMI to participants with clinical depression [92]. Participants received both mood state predictions, and motivational messages when self-reported mood fell below a threshold. Both features were positively received and raised self-awareness.

Reminders

A study of design recommendations for a cystic fibrosis app had participants suggesting that medication reminders be sent in response to self-report of administration [93]. A similar function where patients would receive notifications based on explicit medication reporting was suggested by a healthcare provider in a study by Simons et al.:

> *"I think it would be really useful if somewhere in the app, say when they're...near the end [they receive a message saying] 'You need to put in a request for repeat prescription' "* [94, p. e31]

For self-management, users with chronic conditions valued the possibility of context-sensitive reminders, such as those triggered at a particular location [93, 95]. One study prompted participants to describe their physical activity after levels of intense activity, or extended non-activity, were detected with a smartphone accelerometer [96]. Only prompting assessment when necessary reduces burden on participants. Moreover, from a healthcare perspective, this increased self-awareness could promote positive behaviour [34].

3.2.1.2 Participant Interactions

The very act of self-reporting may increase awareness of behaviours and contextual influences which, in ESM for research purposes, is undesirable given its effect on data validity. In clinical practice, however, raised self-awareness can improve health outcomes by giving patients a sense of independence and knowledge about their conditions. As well as simply having participants complete self-report assessments, smartphone ESM affords new possibilities for participant interaction, that could incite behaviour change, or simply improve study compliance.

Interaction D - Viewing aggregated, collected data

High-definition screens, sophisticated sensing and growing storage capacity of smartphones enable rich visual charts of historical data to be displayed. Research suggests that visual feedback could raise engagement in patients with chronic health conditions, as well as compliance in research participants.

Participants trialling the *MONARCA* system valued both the ability to correlate their moods with implicit data, and to determine the temporal antecedents of low moods [97]. Remote monitoring technology was also welcomed by users with ADHD as a means to track symptoms over time through visual graphs, and identify unusual discrepancies [94]. Users with chronic conditions express interest in viewing their passively sensed information, particularly in the form of visual graphs, supported by participants with mixed chronic conditions [98], diabetes [99], and even serious mental illness [100].

From a research perspective, allowing participants to access visualisations or other means of interpreting their own data could improve compliance. A study by Hsieh et al. showed that providing participants with graphical summaries of their data increased compliance by 23% [58].

Interaction E - Tailoring app to personal preferences

Despite the motivational influence of empowerment to self-manage, maintaining a high level of engagement with apps is still a significant challenge. mHealth apps are easily removed or forgotten about, and patients thus must perceive sufficient value from use if they are to retain engagement. Indeed, recent statistics illustrate that a quarter of apps downloaded are only used once[1]. A balance must be sought between ensuring that app content is based on the input of a professional, while also taking into account the preferences of its end-users. The same appears to be true of psychology ESM apps - when participants are *"researched objects"*, isolated from the researcher and simply expected to comply with rigorous study schedules, this could induce non-compliance [101], and thus undermine the effectiveness of the ESM method.

User-centred design studies of health apps have elicited the importance of flexible prompt schedules to end-users [95, 102]. Moreover, end-users also desire to control the content of feedback that they are provided with [97, 98], as well as the format in which this feedback is presented [94]. For example, participants in a study on health data preferences by Miyamoto et al., although some participants wanted comprehensive information on blood pressure, sleep, and mood, others wanted minimal information such as caloric intake [98]. Similarly, in a study by Simons et al. on technology for monitoring ADHD symptoms, participants desired the presentation of both information input and output to be tailored to their preferences [94].

In research-based ESM applications, many studies begin by manually pre-programming mobile-

[1]http://info.localytics.com/blog/24-of-users-abandon-an-app-after-one-use Accessed 17/03/18

based study applications to participants' sleeping schedules. However, these schedules are subject to change on a daily basis, and flexible smartphone apps could provide participants the ability to adjust these preferences. For example, Markopoulos et al. showed that participants who could choose the time window in which they received assessments showed a significantly higher response rate than participants who received the questions on preset time slots [103]. In their study into human factors of smartphone surveys, Vhaduri and Poellebauer also proposed the use of personalised survey schedules to increase compliance in participants [104].

Interaction F - Providing feedback to researcher

While visual feedback and personalisation are both useful features for improving compliance and engagement, certain issues relevant to study app use may require a direct channel of communication from participant to researcher.

In the medical domain, *"healthcare partnership"* was an emergent theme in one focus group study [98]. Although context-augmented feedback can support patient self-management, focus group participants expressed a desire to contact clinicians for additional explanation of their received feedback. Involving clinicians in this *"sensemaking"* process was proposed to support participants in attaining maximum benefit from ESM. Additionally, individuals with cystic fibrosis identified how self-monitored data would enable them to provide clinicians with feedback between appointments [93].

Research investigating incentive mechanisms for participatory sensing suggests that participants would value being able to provide researchers with feedback. Anawar et al. found that a "participant feedback" incentive mechanic was a prevalent feature of popular weight loss participatory sensing apps [70]. In their participatory design process, Ludwig et al. found that it was important that participants feel like direct contributors to a study. Moreover, this feedback is beneficial from the researchers' perspective - employing participatory design as part of an ESM deployment could enable researchers to address issues that could reduce the effectiveness of the data collection, through direct participant feedback [105].

3.2.1.3 Researchers' roles as developers

Prior to smartphones, the researcher's role in an ESM study was largely passive, and would typically involve installing an application, developed by a professional programmer, onto bespoke devices. These would then be distributed to participants, and collected following the study period for data download and analysis. In contrast, with instant data synchronisation possibilities, researchers can take a more active role in studies, by viewing real-time data and contacting participants if non-compliance is observed. Clinicians could play a similarly active role in mHealth ESM apps, with the ability to monitor patient data and provide feedback. Most importantly, in the context of this thesis, is the potential for researchers to independently create and deploy a study app, and modify it in real-time based on feedback received from participants or from their data.

Researchers could thus act as developers of smartphone-based ESM studies, creating and modifying apps to meet their diverse requirements. Smartphone technology would also afford further activities that support the researcher in this process. A description of researchers' potential activities follows, benefits of which are supported by evidence from literature.

Interaction G - Viewing participant data in real-time

Researchers and clinicians could benefit from viewing incoming participant data during a study, or in between clinical appointments, affording their ability to react to potential issues as they occur. Data analysis could also be begun prior to study completion, in order to inform more useful, effective post-study interviews. For example, in their observations of researchers conducting diary studies, Carter et al. explained:

"Experimenters did not have time to review captured data before elicitation interviews, which curtailed their ability to prepare for the interview and increased the chance that important themes would be missed" [106, p. 126]

In clinical practice, instant access to self-monitored patient data could guide treatment decisions in face-to-face clinical appointments. Study participants living with cystic fibrosis [93] and ADHD [94] both strongly supported the provision of contextual information to clinicians. In an evaluation study of an app where such information access was implemented, paediatricians described how this saved time, focused appointments, and facilitated communication about difficult issues during these appointments [107].

Interaction H - Updating study protocol in real-time

Studies that cite the benefits of smartphone-based ESM all describe bespoke apps developed for the specific purpose of that study. Hence it appears that, although these benefits are generic and adaptable, their implementations are not. Thus, the researcher must be able to adapt these features found in bespoke apps to any ESM study they choose to run. In doing so, they perform end-user development (EUD) activities. (The field of end-user development is discussed further in Section 4.1.)

While theoretical models of behaviour can serve as guidance frameworks for app design, such theories do not capitalise on the full functionality of smartphone devices [108], nor do they account for the variation in requirements of different participant groups. Thus, the involvement of clinicians and patients in the design and evaluation of self-monitoring apps is important for ensuring their sustained use and utility [109]. Allowing researchers to create and modify their own ESM apps is central to the research questions of this thesis; it is hypothesised that, without EUD, the many potential benefits of smartphone ESM are constrained by the availability of a professional programmer to implement a bespoke application. Following such an implementation, any required modifications, however large or small, must be performed by the programmer. The significance of EUD is that the beneficial features of smartphone ESM are put directly into the hands of the researchers. Participants' data and feedback can be directly acted upon, shaping apps in response to their dynamic and diverse requirements.

Interaction I - Providing feedback to participant

Feedback to participants could take a variety of forms, including lifestyle suggestions in clinical practice, or simply requests for compliance in a research study. In either case, there is support for provision of feedback through an ESM app. From their observation that researchers only downloaded data at the end of the study, Carter et al. noted an additional problem of this pattern:

> *"Some of the media captured by users did not align with the goals of the study, and in one case experimenters did not address this until the experiment ended. Being able to view user captures in real time could help with feedback"* [106, p. 126]

From the participatory sensing literature, feedback from researchers could act as a non-monetary incentive mechanism, motivating them as active contributors to a study. A participant in the user-centred design study of Ludwig et al. explained this as follows:

> *"...one has to give the user something in return. So that they know that there has been progress and that based on that, new goals can be developed collaboratively with the user"* [101, p. 498]

In user-centred design studies of healthcare apps, feedback provision from clinicians to patients was also extensively discussed. Given that clinicians have little time outside of scheduled appointments, there was unexpected enthusiasm for this type of interaction. It has been recognised that in-the-moment professional assistance on coping with cancer-related pain, prompted by self-assessment data, would support sustained engagement with self-monitoring [88]. Clinicians also suggested that feedback would not have to be a time-critical intervention, and were enthusiastic about providing regular feedback to patients on their data [110], with such feedback also supported by patient recipients for its ability to enhance understanding [98]. In an implemented trial, clinicians valued a feature allowing response to alerts on patients' symptoms of potential heart failure, suggesting that the initial workload involved in responding to these alerts would be

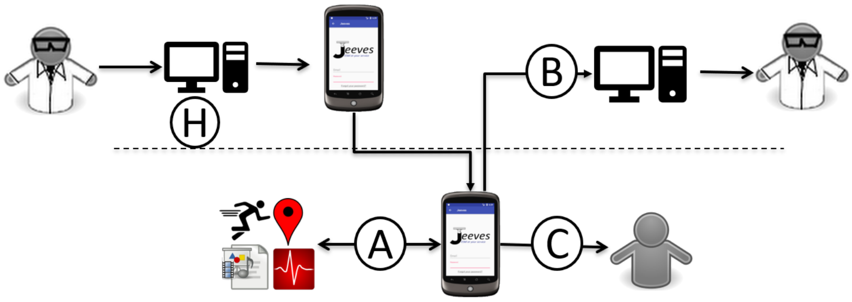


Figure 3.2: A workflow representation of Implicit Context Triggering

reduced in the long-term by minimising unnecessary hospital admissions [89].

From a patient perspective, there may be distrust of automatic context-aware features, particularly in intervention delivery [102], so while context sensing provides opportunities for automatic monitoring and feedback, automation should not replace human input. For example, study participants with diabetes suggested that a 'virtual coach' would only be useful if tailored feedback was provided, and expressed more enthusiasm for human contact [99].

3.2.2 Features of interaction model

From the model of interactions in Figure 3.1, four explicit beneficial features of future smartphone ESM apps are derived, which could be incorporated into the design of tools for developing such apps, namely: **implicit context triggering**, **explicit context triggering**, **two-way feedback**, and **preference tailoring**. The extent to which these features are incorporated into an ESM app deployment is at the discretion of the researcher, directly influenced by their end-user development activities (Figure 3.1.H). The remainder of this section describes the benefits of such features in an EUD tool for ESM apps. Table 3.4 summarises how the features are related to sequences of the interactions previously described, with examples of their application. Further, these interaction sequences are illustrated as workflows extracted from Figure 3.1.

3.2.2.1 Implicit Context Triggering

Automated capture and inference of implicit sensor data (Figure 3.1.A), can support the delivery of assessments, interventions, or reminders at ideal times. Given the range of possible contexts mentioned in the literature, including location, activity, presence of others, noise, and mobile usage, it would be highly beneficial to allow researchers to easily specify the contextual conditions under which notifications should be sent. Further, implicit context could be used to deliver alerts to external contacts such as clinicians or carers. Figure 3.2 illustrates the workflow of such a

Table 3.4: Potential features of ESM apps, and the relevant interactions from Figure 3.1

Benefit	Relevant Interactions	Additional Explanation
Implicit Context Triggering	**H→A→C** **H→A→B**	Researcher defines context and action (H) App detects trigger situation (A) App prompts participant (C) or alerts the researcher/emergency contact (B)
Explicit Context Triggering	**H→E→C** **H→E→B**	Researcher defines context and action (H) Participant inputs self-report (E) App prompts participant (C) or alerts the researcher/emergency contact (B)
Two-Way Feedback	**G→I→F** **B→I→F**	Participant issue is detected through missing data (G) or direct alert from app (B) Researcher contacts participant to resolve issue (I) Participant responds (F)
	D→F→I **C→F→I**	Participant is concerned about their data (D) or receives intervention message (C) Participant contacts researcher to resolve issue (F) Researcher responds (I)
Preference Tailoring	**G→H→E** **F→H→E**	Issue arises through missing data (G) or participant feedback (F) Researcher adds tailoring function to app (H) Participant tailors app to resolve issue (E)

feature, wherein implicitly sensed contexts of interest are defined by a researcher (H) which, upon detection by the app (A) can prompt participants (C) or researchers themselves (B).

3.2.2.2 Explicit Context Triggering

A deliberate distinction has been made between *implicit* and *explicit* context. Implicit context refers specifically to the classification of implicit context such as location, activity, device usage etc. in order to trigger a relevant assessment or reminder. Explicit context, on the other hand, refers specifically to the ability for the device to perform functionality, such as giving relevant feedback to the participant, based on their explicit, self-reported information. The relationship is represented in Table 3.4, and illustrated in Figure 3.3, as E→C and E→B to represent how an ESM app could trigger a response to a participant (C) or a researcher/clinician (B), following a self-report that satisfies the particular trigger conditions.

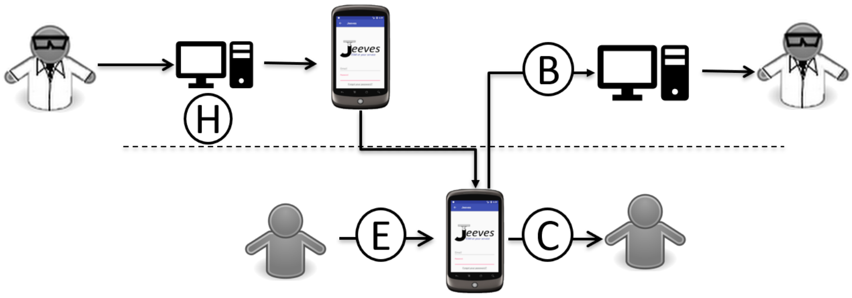


Figure 3.3: A workflow representation of Explicit Context Triggering

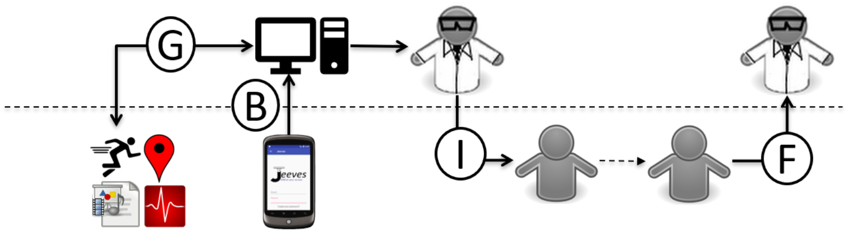


Figure 3.4: A workflow representation of Two-Way Feedback

3.2.2.3 Two-Way Feedback

In clinical applications, automation is not a substitute for a patient-clinician relationship. Similarly, while less common in rigorous study protocols, researchers and participants could benefit from two-way communication to inform acceptable study designs and improve compliance. Table 3.4 briefly summarises examples of interaction sequences that could give rise to two-way feedback, initiated by either researchers or participants.

For example, if a researcher views a participant's data and finds a discrepancy (G), or receives a direct alert via the ESM app (B), they can initiate a dialogue to ensure that the participant is coping with the study, represented as G→I→F and B→I→F respectively. This workflow is also illustrated in Figure 3.4. Likewise, if the participant is concerned about their data (D), or receives an unusual intervention message in response to their self-report (C), they can contact the researcher to seek reassurance.

3.2.2.4 Preference Tailoring

A major problem in ESM studies, on paper, smartphones, or otherwise, is non-compliance and attrition resulting from the high burden of completing multiple surveys. Notification reminders can be sent to prompt compliance, but when these notifications are deemed intrusive in participants' lives, this can aggravate the issue further. Participant attrition consequently causes burden for researchers, as the statistical power of study results is dependent on a larger sample

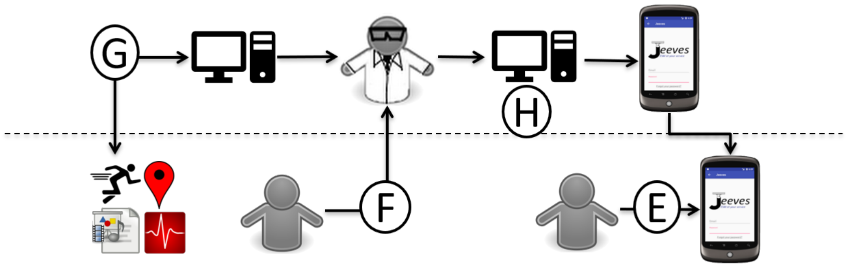


Figure 3.5: A workflow representation of Preference Tailoring

size. Attrition could also be a particular concern in the medical domain, where patients' health outcomes could be directly related to their compliance.

Hence, as expected, user-centred design studies report a need to tailor assessment timing and questions to participants' personal characteristics. Again, Table 3.4 shows how this feature relates to possible interactions in the model. For example, a researcher may view participants' data and find that assessments are not being completed at certain times or under particular conditions (G). They could resolve this by providing participants with the ability to tailor these times to their own preferences to maximise the chance of completion, represented by the G→H→E interaction sequence. Likewise, a participant could proactively contact their researcher to request more flexibility in survey scheduling if they are struggling with inconvenient sampling times, represented by F→H→E. These interaction workflows are also illustrated in Figure 3.5.

3.3 Design Implications

The model of interactions proposed in this chapter synthesises results from various user-centred and participatory design studies, guidelines for effective mHealth interventions, and studies investigating factors relating to ESM study compliance. The design features implemented in Jeeves, described in Chapter 5, are derived from the implications of this model, which are prescriptive interpretations of the features described in Section 3.2.2.

Firstly, in relation to the issue of participant burden highlighted in Chapter 2, **implicit data collection from smartphone sensors should be incorporated**. Foremost, this alleviates the need for participants to enter this data themselves, but can also be used to infer appropriate situations in which to prompt participants.

The second design implication is that **prompts based on participants' unique responses should be incorporated**. This relates to the issue of researcher and clinician burden raised in Chapter 2. For researchers running ESM studies with large samples, this could include automatic prompts based on a participant's lack of response. For clinicians monitoring patients, this could allow messages of support or information to be sent to at-risk patients at any time.

While these two implications stem from situations where there is a need to minimise direct interaction from either participant/patient or researcher/clinician, the studies reviewed highlight a frequent need for direct human intervention, motivating the third implication that **a means of direct feedback between researcher and participant should be incorporated**.

Finally, Chapter 2 noted the diversity of participant samples and the consequent variability of study specifications. Extending this *between*-sample variation, an issue raised in this chapter is the need to address individual preferences, or *within*-sample variation. Thus, the fourth implication is that **a means for participants to tailor apps to their personal preferences should be incorporated**.

With the exception of the need for two-way feedback, providing researchers and clinicians with the means to create and adapt ESM study protocols is central to the aforementioned implications.

3.4 Chapter Summary

This chapter presented a literature review of developments in smartphone-based data collection studies, and a further review of studies wherein desirable features of such smartphone apps were obtained from stakeholders. Derived from these reviews, implications for the design of future ESM app-based tools are proposed. Thus, the purpose of this chapter is to provide an answer to the research sub-question: *"What are the necessary features of apps that facilitate remote assessment of research participants or clinical patients in their natural environments?"*. Through triangulating this literature review with interview and case study feedback in Chapters 7 and 8, the model is refined and an indication of its external validity is given.

As discussed, the utility of the features derived from the interaction model is contingent on the researchers' power to define the specific details of these interactions, without the need for programming. Thus, the question still remains of how to allow a researcher to define and moderate these interactions, by successfully implementing them into a usable EUD tool for ESM app creation. In order to address this question, the next chapter enters into the field of end-user development, and effective paradigms for use by non-programmers.

CHAPTER FOUR

END-USER DEVELOPMENT

At the core of this thesis is the question of how to enable the development of experience sampling smartphone apps, which satisfy the diverse requirements of all stakeholders, without the reliance on an explicit step of professional programming. Thus, this chapter focuses on the concept of end-user development (EUD), framing this work within human-computer interaction.

One pertinent aspect of EUD is the development paradigm, the range of which will be surveyed in this chapter, with a specific focus on visual programming. In addition to the techniques used to support end-users' mental models of programming, EUD introduces socio-technical issues regarding its adoption into working practices, which will also be discussed.

EUD in the specific domain of ESM has not been widely applied. A survey of EUD tools that have been developed to facilitate the creation of ESM studies represents a main contribution of this chapter. These two reviews serve to present the requirements of an ideal EUD-ESM tool, the current state-of-the-art, and thus the discrepancy between the two - a foundation in the design and development of Jeeves.

4.1 EUD and Visual Programming

"EUD can be defined as a set of methods, techniques, and tools that allow users of software systems, who are acting as non-professional software developers, at some point to create, modify, or extend a software artifact." [16, p. 2]

This widely quoted definition of EUD comes from the work of Lieberman et al. in their seminal publication describing the field and its benefits. As an area of research, EUD has grown considerably since its conceptualisation, and indeed it is no longer the “emerging paradigm” previously described, with a growing number of computer users performing EUD activities, whether they are aware or not. In 2005, Scaffidi et al. estimated that in the USA alone, there are over 12 million people who claim that they “do programming” as part of their job, and the number of software end-users outnumber professional programmers by thirty to one [111]. EUD has thus grown out of recognition that professional software developers can not possibly anticipate the diverse needs of these end-users in advance.

Tools allowing end-users to create and modify software are equally diverse in their complexity, domain, and motivation. Kelleher and Pausch divide their review of EUD tools by two primary motivations: programming *for its own sake*, and *to accomplish a specific task* [112]. The former category includes interesting examples of environments designed to scaffold novice programmers in their goal of learning a high-level textual programming language, such as Java or C++.

Conversely, the latter category encompasses EUD tools for whom software development is a means to an end, rather than the ultimate goal. For such “professional end-user developers”, often a key challenge is not one of education, but of infrastructure, and how such EUD environments can be incorporated into current working practices [113]. As such, as well as the two categories of end-user developer, this chapter distinguishes two research areas of EUD:

1. **Programming** - The end-user’s interaction with the EUD tool
2. **Organisation** - The EUD tool’s interaction with the end-user’s environment

The first considers the methods, tools and techniques of EUD with regards to how well they capture end-users’ mental models of programming, and how to maximise the usability and flexibility of such approaches. Research methods in this area are often lab-based usability studies with specific tasks designed to test novel features of a particular tool.

The second acknowledges that the unit of study is not just the individual, but how an EUD tool would evolve in working practice across communities and organisations. An issue often not addressed by lab-based usability evaluations of EUD tools is how they would realistically be adopted in practice, such that research must move into the field with case studies and field studies. This issue considers EUD as more than just a *“set of methods, tools and techniques”* but as, potentially, a complete reinvention of working practices and cultures. The following section first addresses **programming** research, describing different levels of complexity involved in the creation and modification process, as well as the programming paradigms employed.

4.1.1 Levels of programming complexity

End-user programming has been referred to as "tailoring" - broadly defined as an activity to modify a computer application within its context of use [114]. Through tailoring activities, end-users can become masters of their own software tools, modifying these tools to suit their own requirements without the need to rely on a professional programmer.

Mørch defines the complexity of tailoring approaches into three categories: **customisation**, **integration**, and **extension** [115]. These categories overlap to some extent with EUD activities described by Lieberman et al., and indeed it is not always possible to place activities into one of the three categories. Instead, they appear to represent three points on a spectrum of complexity.

4.1.1.1 Customisation

Mørch's definition of customisation is similar to the definition employed by Lieberman et al. of *"parameterization and customization"* [16]. Such activities involve the adjustment of parameters that allow an end-user to change the content of an application, but not its behaviour. "End-user authoring" tools, such as those described for conducting participatory sensing campaigns in Section 3.1.1, fall into this category through supporting content modification while leaving the programmatic structure intact. For example, they allow users to change the design of a questionnaire form, but not the conditions under which it is presented. Another customisation activity discussed by Lieberman et al. is that of "annotation" - supplementing programs, data and results with comments, as a form of secondary notation to impart additional meaning.

This thesis argues that, although customisation is clearly a tailoring activity, it is not end-user development; adjusting parameters of an existing application's behaviour is not the same as adjusting the application behaviour itself. As a simple example, adjusting the time on an alarm clock application would not be considered as development. However, development could be integrating a function that snoozes the alarm for a number of minutes, and it is this *integration* process that forms the basis of many EUD tools.

4.1.1.2 Integration

Integration is closely related to the definition of *"program creation and modification"* by Lieberman et al., who also consider integration activities to be relevant examples of EUD, as opposed to customisation [16]. Conceptually, end-users compose existing program functionality to define their chosen behaviour. While no modification of source code is required, integration activities can enable a wide range of program behaviour through linking and swapping of high-level components.

Program modification can also involve *"extended parameterisation"*, whereby end-users select and integrate functionality created by other end-users, stored in a shared repository [16]. However, such activities still only allow for combinations of existing functions. Creating new functions requires modification beyond the scope of integration, which can instead be accomplished through *extension* activities.

4.1.1.3 Extension

Extension is the most complex manifestation of EUD, whereby functionality exists to allow end-users to define their own components, as opposed to simply integrating pre-existing ones. This level of complexity generally requires the end-user to perform some modification of source code to realise these components, and thus bridges the gap between end-user programming and *real* programming.

Mørch's view of the end-user in such development activities is in line with that of Segal, who distinguishes "professional end-user developers" as those for whom software development is already part of their working practices to some degree [113]. Indeed, Spahn et al. consider such users to be professional programmers [116], transcending their role as domain expert developers.

4.1.1.4 Roles of end-user developers

In contrast to the often steep learning curves of professional programming languages and development environments, EUD tools should offer a low threshold to entry. In addition, dependent on the desired flexibility of development enabled, tools could support a user from simple customisations to formal scripting of applications, through the three levels of tailoring previously described. As such, EUD tools have the potential to provide a smooth transition into real programming, or simply to provide sophisticated modification abilities without the need to ever learn such real programming.

In their recommendations for EUD design principles, Spahn et al. advocate a *"gentle slope of complexity"*, allowing end-users to move seamlessly from simple customisation activities to more complex extensions [116]. Additionally, they also define three types of end-user by borrowing Nardi's definitions of *non-programmer*, *local developer* and *programmer* [117]. In this definition, *local developer* refers to any end-user engaged in EUD practices, but not proficient in high-level textual programming. Their thesis is that appropriate design principles could allow end-users to start as non-programmers and become programmers, but within a single embedded EUD environment, as illustrated in Figure 4.1. This figure also illustrates their suggestions of testing, community, and appropriation support for scaffolding end-users through their increasingly complex development activities.

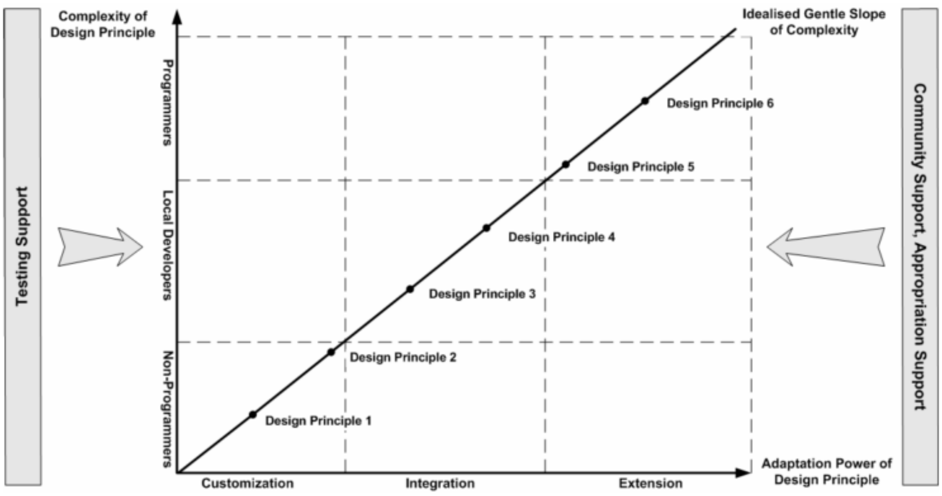


Figure 4.1: Gentle slope of complexity from Spahn et al. [116]

An alternative perspective is that of the Software Shaping Workshops (SSWs) methodology [118]. This concept places design stakeholders in three separate roles: end-users, domain experts, and meta-designers. Each role has a corresponding SSW, conceptualised as an "artisan workshop" in which each user group finds the tools, and *only* the tools that they need to perform their relevant activities. As such, rather than considering the end-user as having a dynamic, evolving role, each end-user has their own workshop solely for performing customisation, integration or extension as appropriate to this role. The concept is illustrated in Figure 4.2, and will be explained in further detail in Section 4.1.4 on organisational factors of EUD.

4.1.2 Programming paradigms

The programming paradigms used in EUD tools are closely tied to the context of use, as well as the complexity of development required. In customisation activities, there is no paradigm *per se*, as the essence of programming is not captured in simple adjustment of parameters. Likewise, in extension activities, the modification of source code is required, such that high-level, textual programming may occur with no intermediate means of representation. Integration, however, is the middle-ground in which domain experts are required to create and modify programs, while remaining abstracted from the complexities of professional programming languages. The software developed in this thesis utilises a visual programming paradigm, which constitutes most of the following related work. Alternative programming strategies from EUD literature are discussed as follows.

Programming By Demonstration (PBD) environments allow the creation of applications without an intermediate representation, wherein a user simply demonstrates the required output

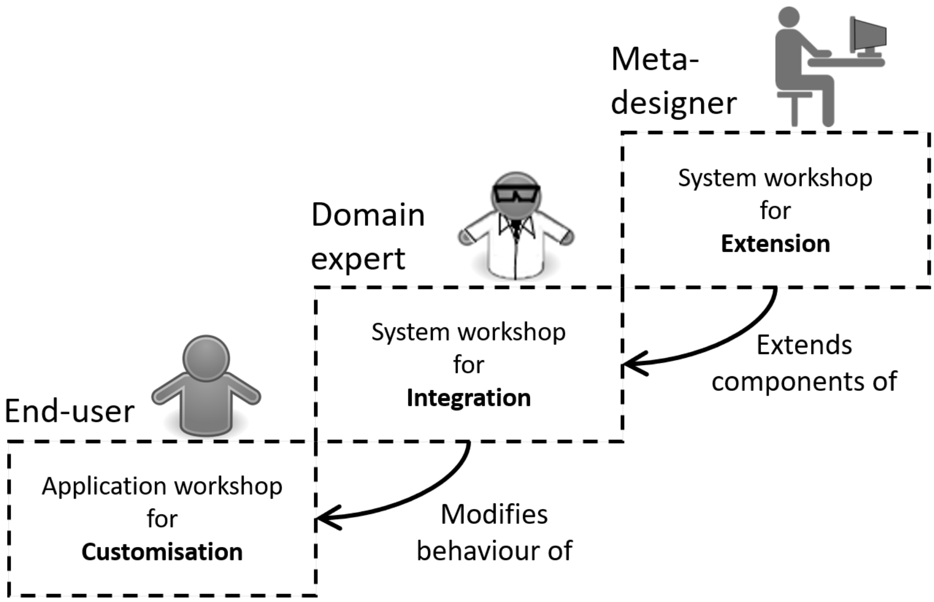


Figure 4.2: Diagram illustrating the hierarchical software-shaping workshops (SSWs) of end-users, domain experts and meta-designers

of an application, and the specification is inferred and automatically generated [119]. PBD thus liberates end-users from having to learn any form of programming semantics. As an example, *a CAPpella* by Dey et al. supports PBD of context-aware applications [120]. In this EUD environment, an end-user developer begins a recording process, demonstrates their context-aware behaviour, and later annotates the recorded actions to allow *a CAPpella* to accurately infer the behaviour required.

Natural programming is a paradigm with a similar goal to PBD, of making the development process as intuitive as possible for the developer. A notable example is the *HANDS* environment by Pane and Myers [121]. In designing the environment, the authors conducted studies into how both children and adults conceptualise programming problems in natural language. They discovered that end-users naturally use event-driven statements, and designed their language as such, with programmatic syntax as close to natural language as possible. In similar work, Liu and Lieberman defined the concept of programmatic semantics as *"a mapping between natural linguistic structures and basic programming language structure"* [122, p. 1597]. Using these semantics, their *Metafor* tool allowed identified parts of speech in a natural language statement to be automatically converted into program constructs.

Programming by specification (PBS) also bears similarities to PBD, in that the specification is automatically generated. However, rather than demonstrating the desired *output*, the user describes the ideal *properties* of an application. Burnett and Scaffidi define PBS environments as those using either visual forms-based interfaces, or very high-level “natural” programming with text [123]. However, these two paradigms are very different from the end-user’s perspective. As such, this thesis considers PBS systems to be only those that allow end-users to tailor applications through web forms and similar visual interfaces, with very high-level textual input considered as a derivative of natural programming interfaces, described in this section. One such example is an environment for e-Government services described by Fogli and Provenza, where developers use a series of forms to automatically generate XML-based web services [124]. In their review of tools supporting event-condition-action (ECA) rules for task automation tools, Desolda et al. observed that the majority of these tools used wizard-based interfaces for guiding non-technical users through the process [125]. Furthermore, they explain how this approach limits users from constructing more complex ECA rules, and that a middle-ground must be sought between development complexity and flexibility.

Finally, **visual programming** is distinct from PBS in that, rather than entering text into a graphical form, the specification of a program involves manipulating the graphical objects themselves. Although this drag-and-drop direct manipulation is conceptually intuitive, visual programming environments are not exclusive to novice users. Indeed, complex visual programming environments exist that, while alleviating the need for textual programming, still require extensive training to be used appropriately. A prominent example is National Instrument’s *LabVIEW*, used in industry by professional scientists and engineers [126]. Although it is a powerful tool in its specialised domain, the extensive documentation, tutorials and training demonstrations available are indicative of its high complexity[1]. As such, visual programming is not constrained within the bounds of end-user development, but exists as a programming paradigm in its own right, defined as any representation of a program that uses two or more dimensions to convey semantic meaning [127].

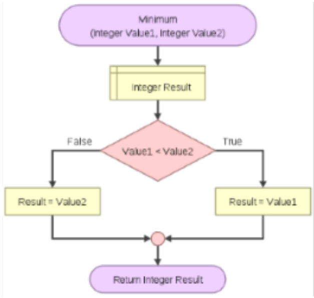
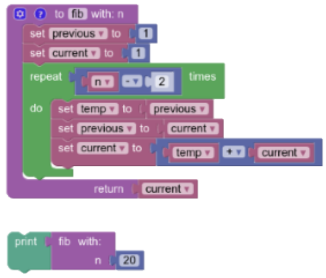
4.1.3 Evolution of visual programming

Visual programming grew in popularity in the early 1990s, with ambitious visions of what could be accomplished by abstracting away from the complexities of writing code. It was perceived that a visual representation would map more closely to an end-user’s mental model of a programming problem. Nevertheless, empirical evidence comparing visual programming languages against textual equivalents was sparse and contradictory [128]. Moreover, attempting to scale visual programming up to incorporate the dense syntax of a general-purpose programming language was an intractable problem, particularly given the constraints on screen space.

In recent years, however, visual programming environments have become more prevalent as

[1]http://www.ni.com/en-gb/shop/labview.html Accessed 17/03/18

Table 4.1: Visual programming paradigms, their formal classifications, and illustrative examples

Paradigm	Diagrams	Icons	Blocks
Class	Connection-based	Geometric-based	Connection-based
Relation	Linking	Iconic sentence	Overlapping
Example			

educational tools for teaching programming syntax and semantics. For children and young adults learning computer science, visual programming environments such as *Scratch* [129], *Alice* [130], and *App Inventor* [131] are appealing both visually, but also in allowing interesting applications to be built with minimal initial effort. Such environments are currently being used in school curricula to teach programming concepts to children and young adults.

Visual programming environments are more than just teaching tools, however, and have proven useful for domain-specific applications, where problem concepts from a limited domain vocabulary can be mapped to visual components, without becoming intractably large. Domain experts can use visual programming in their work practices without the need to learn textual syntax and semantics. The aforementioned *LabVIEW* is a prime example [126], as well as Max, a visual language for audio processing applications[2].

Visual programming is abstractly defined as programming in two dimensions, covering a wide variety of concrete implementations. Previous work from the 1990s described taxonomies of visual programming languages [132, 133]. However, in the two decades since these classification systems were published, a wealth of new visual languages have been developed. A more recent classification system was proposed by Costagliola et al. [134], which can be used to classify the visual metaphors used in modern environments. These metaphors tend to fall within one of three categories: diagrammatic programming, iconic programming, and blocks-based programming. Examples of these metaphors with classification details are illustrated in Table 4.1. Each has its own benefits and drawbacks, which will be discussed as follows.

[2]https://cycling74.com/products/max/ Accessed 17/03/18

4.1.3.1 Diagrammatic programming

Diagrammatic programming is one form of a "connection" visual language as defined by Costagliola et al [134]. Diagrammatic languages resemble flowcharts, wherein programming constructs are represented as boxes, and connected together in various ways to visually represent control flow. A prominent example of this in the educational domain is *Raptor* [135]. *Flowgorithm*[3] and *Visual Logic*[4] are other commercial applications that can be used to create executable flowcharts. The aforementioned professional environments *Max* and *LabVIEW* use a data flow diagrammatic approach to allow the developer to 'wire' components together. As a paradigm employed in task automation tools, Desolda et al. implemented a diagrammatic programming paradigm in their comparative usability study. However, it emerged that even study participants with technical expertise found this paradigm to be the worst in terms of performance and satisfaction over PBS-based alternatives. They concluded that *"even technical users, who should be acquainted with wired notations, perform better and are more satisfied if using other interaction paradigms"* [125].

4.1.3.2 Iconic programming

Iconic programming languages fall within the "geometric" visual language class. In iconic programming environments, code is represented by graphical symbols (or icons) that can be spatially arranged to represent how an application should function. Geometric languages are distinct from connection-based visual languages in that there are no explicit connections between the icons. Instead, the spatial location of icons relative to one another determine their relationships. One example is the *Kodu* environment, which uses graphical icons to allow developers to compose program behaviour in the form of event-driven clauses [136]. *GALLAG Strip* is another similar environment that represent actions and responses as graphical icons, for composing rules in context-aware applications [137].

4.1.3.3 Blocks-based programming

A blocks-based programming language is formed of constructs resembling jigsaw puzzle pieces. Like diagrammatic languages, they are also "connection" visual languages, wherein the visual elements have insets and protrusions that show how they should fit together in the correct way, consistent with the jigsaw puzzle metaphor. This visual syntax reduces the possibility of errors and allows the user to focus on the semantics of their application [138].

[3]http://www.flowgorithm.org/ Accessed 17/03/18
[4]http://www.visuallogic.org/ Accessed 17/03/18

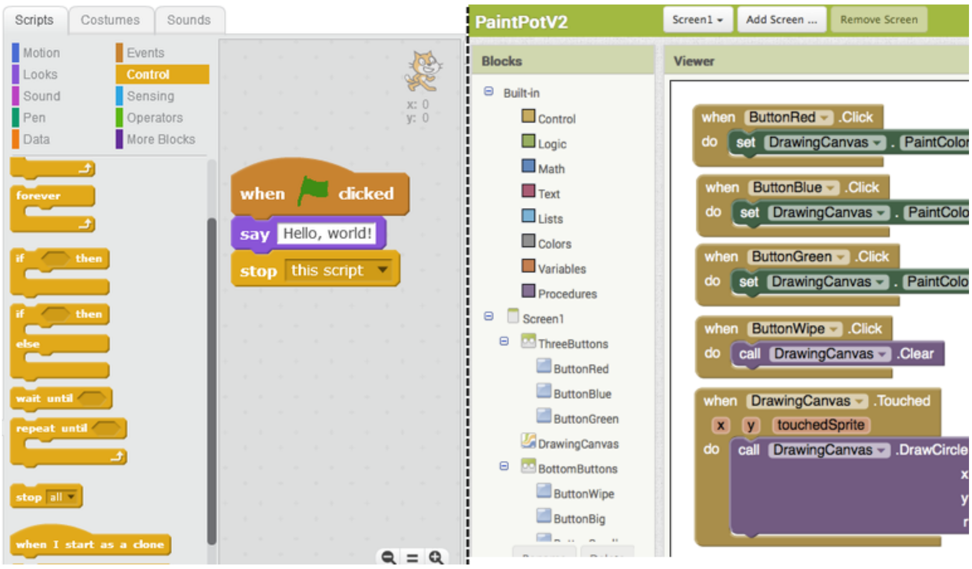


Figure 4.3: Palettes and canvas of Scratch (left) and App Inventor (right)

Scratch is a prominent example of blocks-based programming [129]. It is targeted at children and young adults, allowing them to drag and drop blocks to create visually appealing games and animations. *App Inventor* is a similar project to the work in this thesis in terms of programming paradigm and development target [131], which allows users with no programming experience to develop a wide range of fully functional apps for Android devices. Both environments are notable for their ability to motivate and engage users, through the rapid development of meaningful applications, but also comprise a variety of programmatic constructs such as loops, conditional statements, and variables that allow complex behaviour to be implemented if desired.

App Inventor utilises *Blockly*[5], an open-source library for creating blocks-based languages. Blockly has supported the creation of a variety of such languages in a growing number of domains, but particularly for educational purposes. A recent article in *Communications of the ACM* provides a thorough discussion of the success of blocks-based programming and the reasons for its utility as an educational tool [138]. While its benefit for supporting students' transition to textual programming is not of relevance to the work in this thesis, the learnability of the paradigm itself is particularly important if it is to be adopted in time-constrained practices.

Blocks-based programming is the development paradigm of choice for Jeeves, the visual programming environment implemented in the work of this thesis. (The specific design details of Jeeves will be discussed in Chapter 5.) From a general perspective, salient features common to all blocks-based environments have been applied, which can be observed in the screenshots of *Scratch* and *App Inventor* in Figure 4.3. These features include:

[5]https://developers.google.com/blockly/ Accessed 17/03/18

- **Drag-and-drop canvas:** In blocks-based environments, the canvas is a central feature onto which developers can drag, drop and rearrange their application blocks. These canvases can be panned and zoomed, allowing users to layout their applications in a flexible way.

- **Categorical palettes:** To support *recognition over recall* [138], all blocks can be accessed through named palettes adjacent to the canvas. Clicking on a palette name displays all the blocks in that category, allowing more experienced users to efficiently find the blocks they need, while minimising effort required for novices to do so.

- **Effective visual metaphor:** The jigsaw puzzle metaphor has been identified as an effective visual metaphor for novice programmers. Program constructs that can not be combined in a syntactically correct way do not fit together as puzzle pieces, thus eliminating syntax errors, which are a major source of frustration in learning to program textually. Bau et al. refer to this feature as *"syntax-directed editing with constrained direct manipulation"* [138].

- **Secondary notation:** Blocks-based languages make substantial use of secondary notation to convey additional meaning to both developers and readers of specifications. The shapes, colours and sizes of blocks are all suggestive of their relationship to one another. Further, the position of blocks in 2D space can be used as an extra form of secondary notation, in contrast with iconic programming languages where 2D position is enforced. Further, blocks do not have to be connected in a network as in diagrammatic languages, which minimises the potential for confusing masses of connections.

4.1.4 Organisational factors

Having considered the challenges to individuals at the programming level, this section discusses a broader unit of study - the organisational and social context. Developing EUD tools that support users in their current practice, and evaluating the success of these tools in doing so, are major challenges, which require investigation beyond participants interacting with a tool in a lab study. In real-world deployments, individuals' interactions with technology are highly dependent on other individuals who form part of their working practices, and existing technologies already used in this environment. These factors are diverse and dynamic, such that continuous communication and collaboration are necessary to ensure that deployed technology evolves with an environment and the people within it. Such requirements can be addressed with a "meta-design" approach [139].

Meta-design, as applied to EUD, is a framework through which EUD tools are designed for evolution and growth by communication, collaboration, and mutual development with the various stakeholders of the tool [118]. Distinguished from user-centred design, meta-design allows design **during** use, as opposed to design **before** use, acknowledging that the requirements of end-users are diverse and dynamic. Fischer et al. explain this as follows:

"*[user-centred and participatory design] force all the design intelligence to the earliest part of the design process, when everyone knows the least about what is really needed*" [118, p. 65].

The Software Shaping Workshops (SSWs) methodology discussed earlier, and illustrated in Figure 4.2, has each stakeholder in software design, from the end-user to the meta-designer, employing their own SSW to shape the development of software using activities appropriate to their needs and abilities [140]. In the context of meta-design, users are not isolated in their development activities and instead make individual contributions through their SSW that are communicated to other users. In doing so, the entire system is a product of iterative design, development, evaluation and feedback by meta-designers, domain experts, and end-users.

This explicit distinction between the end-user and the domain-expert challenges the more conservative view of EUD that developers are typically creating software for *personal*, rather than *public* use [141]. Indeed, activities span a wide range of purposes, from private EUD (development for personal use) to public EUD (development for a community of users). Cabitza et al. define public EUD as the process of non-programmers developing applications for *other people* to use [142]. A further distinction is made in their work between public-*inward* EUD - developing for a group of users within the same community or organisation, and public-*outward* EUD - developing for a different community of users. The latter represents the work in this thesis - Jeeves is an EUD tool for researchers and clinicians to develop apps for their participants and patients. With end-users, domain experts, and meta-designers working independent of one another, but towards a common goal of useful, usable software, the meta-design approach could resolve these challenges that transcend individual EUD.

4.1.5 EUD evaluation

Given the contextual influences of EUD in practice, it is surprising that a prevalence of lab-based usability studies in the evaluation of EUD is contrasted by the lack of research into real-world utility [18], a disparity recognised within HCI as a whole [143]. EUD evaluations are largely focused on the programming paradigm and how users' mental models of programming tasks affect the usability of particular paradigms. However, the development of *useful* EUD tools requires knowledge of who potential end-users are, what their goals and motivations are, and how such tools could fit with current working practices. Tetteroo and Markopoulos recently conducted a review of research methods in EUD, relating the methods used to the research purpose, such as

Table 4.2: Summary of review of EUD research methods by Tetteroo and Markopoulos [18]. Numbers indicate how many of 93 papers have the corresponding research purpose and setting

Setting	Understanding	Engineering	Re-engineering	Evaluating	Describing
Natural	2	0	0	8	11
Artificial	**13**	0	1	**25**	4
Independent	2	42	6	6	15

describing an ideal tool, or evaluating an existing tool [18]. A numerical summary of their results is shown in Table 4.2, which presents how many of the 93 papers they classified have a particular research setting, and research purpose. Two results of particular interest are highlighted in this table, which shows that the majority of research into understanding and evaluating EUD tools has been in lab settings, rather than in the field or through qualitative interview and survey feedback, for example.

These results are surprising because, as previously stated, evaluating the success of an EUD tool in its envisioned context of use gives more insight and direction to future EUD use than a lab-based usability study. Moreover, *understanding* the problems faced by end-users seems best suited to field work, where these problems can be observed first-hand, or to survey and interview research, which elicit first-person feedback from end-users themselves, independent of setting. One reason for this discrepancy could be that lab-based studies are simply cheaper and less time-consuming to conduct. They also provide quantitative results that are more straightforward to analyse, as opposed to often complex and disjointed qualitative data from field methods.

Despite this discrepancy, some recent work has focused on studying users in a qualitative way to seek understanding of current practices, and evaluations of systems in situ. For example, Namoun et al. discuss design implications for mobile EUD from results of surveys and focus groups [144], in which they investigated factors influencing mobile users to create apps for themselves in their spare time. Although this is *individual EUD* as defined in the previous section, and is not applied within a meta-design framework, the qualitative methods used and subsequent data analysis provided a model of mobile users' intention to perform EUD activities

In work outside of EUD, models have been derived that predict the adoption of general technology, including the Technology Acceptance Model (TAM) (illustrated in Figure 4.4), and the Unified Theory of Acceptance and Use of Technology (UTAUT) [146]. Core principles of both these models are that a user's intention to adopt a technology is strongly influenced by both usability,

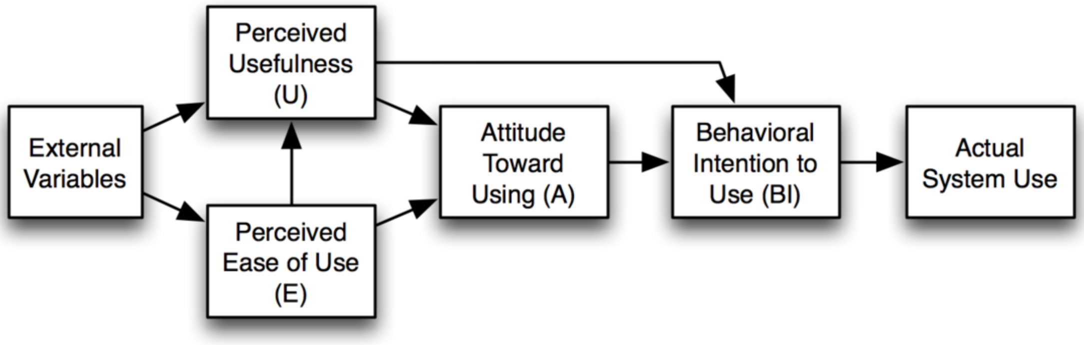


Figure 4.4: The Technology Acceptance Model [145]

and usefulness. While these models have been applied across a variety of software, EUD tools are unusual, in that they may introduce a meta-design approach, where tools and developed artifacts are dynamic and involve end-users in a non-traditional role as developers. Thus, the overall research question "*What factors influence the adoption of technology for researchers and clinicians to develop experience sampling smartphone apps?*" is focused on EUD in a highly specific context, but aims to generalise to a model of EUD acceptance.

4.2 End-User Development of ESM Studies

The review of recent ESM studies in Chapter 2 highlighted a small number of existing systems that researchers have used to create their applications. In order to acquire a more comprehensive list of tools facilitating creation of ESM apps, a further literature review was conducted. Computer science literature databases including the ACM, IEEE and Scopus digital libraries were searched using the terms 'experience sampling', 'ecological momentary', 'end-user development', and 'end-user programming', to derive tools developed in research. The related work of each of the selected publications was also examined to ensure that similar referenced examples were not omitted. Google Scholar was then used to find forward-citing papers, in order to determine whether any empirical studies that utilise the cited tool were available. Given that some tools exist in the commercial domain, a standard Google search was used with the search terms listed above, in an attempt to find proprietary examples. Finally, Conner's resource on ESM creation tools was used to identify those satisfying the criteria described as follows [147]. (For the sake of brevity, EUD tools for ESM app creation will be hereby referred to as *EUD-ESM tools*.)

Exclusion criteria were applied to limit tools to those relevant to the work of this thesis. For example, those not intended for use by non-programmers were excluded, such as Apple's *ResearchKit*[6]. Although it is highly flexible and supportive of health app development, researchers are required to have knowledge of Swift or Objective-C programming in order to apply it. Further, many tools only support survey creation, with no prompting mechanism. For example, *Qualtrics*[7] provides an interface for creating online surveys and analysis of survey results. However, with no means to prompt participants automatically (researchers must send links to online surveys in SMS messages) it cannot be considered an EUD-ESM tool. Similarly, tools that rely on participants to self-initiate assessments were excluded. Examples include the aforementioned participatory sensing platforms such as *Open Data Kit* and *Sensr*. Further details of tools were acquired by reviewing relevant literature, and by testing them when available. For those that could not be tested, authors of publications describing the tool, or the software developers in the case of commercial tools, were contacted to request a trial or further details. If this was not possible and no further information could be obtained without payment, the tool was not included. For example, *MetricWire*[8] and *mEMA*[9] were not included for this reason. In some cases, researchers' provided information was also sufficient to exclude the tool. The remainder of this section provides details on the final selection with regards to their strengths and weaknesses, and in particular to their support of useful ESM app features derived in the previous chapter.

4.2.1 Pioneering tools

Designed for bespoke devices or obsolete mobile phones, these tools were not included in the table of those currently available for smartphone-based ESM. Nevertheless, their contributions are significant in inspiring the development of the more recent tools in this table, and are thus described.

4.2.1.1 ESP (2000)

ESP (Experience Sampling Program) is the earliest example of ESM creation software that was discovered in this review [46]. Applications created with ESP are runnable on PDAs or similar bespoke study devices. As described in its documentation, running a study using ESP involves distributing PDAs instrumented with the software, and collecting these devices back at the end of the study period for data upload and analysis. The created ESM app prevents other applications from running, meaning that the PDA serves no other useful function to participants. While some advanced survey features are available, including probabilities of

[6]https://developer.apple.com/researchkit/ Accessed 17/03/18
[7]https://www.qualtrics.com/uk/ Accessed 17/03/18
[8]https://www.metricwire.com/ Accessed 17/03/18
[9]http://mema.ilumivu.com/ Accessed 17/03/18

Table 4.3: Pioneering platforms and their distinguishing features

Tool Name	Year	Devices used	Configuration	Data transfer	Instant feedback
ESP	2000	Palmtops	CSV text file	None	No
CAES	2003	Pocket PCs	CSV text file	None	No
MyExperience	2007	Windows Mobile	XML file + scripting	Wireless - Opportunistic	No
Momento	2007	Pocket PCs	Text file + GUI	SMS / MMS	Yes

questions appearing, and conditional branching, all question details including this advanced logic have to be defined in a text editor, making implementation more cumbersome, and affording syntax errors. Interval-contingent, signal-contingent and event-contingent studies are possible. However, mixed sampling methods are not, and each study must adhere to only one of these sampling schedules.

Despite its many limitations, as an open-source platform ESP acted as a scaffold upon which more advanced functionality was built in further research. Moreover, as the first example of an EUD-ESM tool, it proved the utility of such software for non-programming researchers.

4.2.1.2 CAES (2003)

CAES (Context Aware Experience Sampling) is a tool developed by Intille et al. which was pioneering in its support for context-sensitive ESM studies. *CAES* is developed as an extension to *ESP* in order to objectively identify the activities of participants outside of a lab environment, free of self-report bias [148]. As sensors are not built-in to PDA devices, *CAES* communicates with external sensors such as a clip-on GPS tracker and tri-axial accelerometer. It was also the first tool that allowed participants to record audio or take a photograph to annotate collected data, provided that the correct PDA hardware plugin was installed.

Like *ESP*, however, study specification are created and modified through a comma-separated text file. Further, *CAES* has the same additional disadvantages inherent in the available technology of the time; PDAs are bulky (aggravated by the need to attach additional hardware for sensing capabilities), and a lack of wireless transfer capabilities means that the software has to be installed manually onto each device, and collected back in-person at the end of a study. This means that compliance can not be monitored in real-time.

4.2.1.3 MyExperience (2007)

As personal mobile devices became more commonplace, the developers of *CAES* contributed to the *MyExperience* project [149], the earliest tool identified in this review that enabled researchers to couple studies to participants' mobile devices. This eliminates the burden on participants of carrying an additional device. Further, *MyExperience* was the first platform to enable opportunistic wireless data transfer of collected data, and also enabled a wide variety of sensor-based trigger conditions.

Again, a primary barrier to *MyExperience* is the means by which researchers configure studies. According to [149], specifications are modified through an XML interface, and further high-level scripting is required to define the conditions under which context-contingent triggers should fire. Nevertheless, it remains a seminal platform in the domain of EUD-ESM, and is still used in recent studies.

4.2.1.4 Momento (2007)

The aforementioned pioneering platforms all require configuration through a CSV or XML text file which, while not programming in itself, is still time-consuming and prone to syntax errors by researchers with little programming experience. *Momento* was the first EUD-ESM tool to provide a desktop-based GUI for researchers to configure study details [106]. It was also the first such platform to support instant two-way feedback between participant and researcher, via SMS and MMS messages. Additionally, the desktop platform of *Momento* enables live monitoring of a study and its communications, previously impossible with older platforms.

The disadvantages of *Momento* over more modern approaches include the need to install and configure the software onto each participants' device manually. Further, as all assessments and feedback have to be communicated through SMS and MMS messages, such studies are potentially costly to participants if run for an extended period of time.

4.2.2 Modern tools

A number of EUD-ESM tools for smartphone devices are currently available, either as free tools for research purposes, or as commercial platforms requiring payment to use. This section first describes the tools created for research purposes, reported in academic publications and generally freely available to use and adapt. While these tools lack the polished appearance and technical support of those in the commercial domain, they offer the researcher more flexibility and eliminate the need to rely on a mediating developer or company. Following description of these research tools, proprietary examples are similarly described. In doing so, the research

Table 4.4: Features derived from Section 3.2 and their implementation in EUD-ESM tools

Tool Name	Implicit Context	Explicit Context	Preference Tailoring	Two-Way Feedback	Creation Interface
iPromptU[10]	🔴	🔴	🔴	🔴	PBS
PsychLog [150]	🔴	🔴	🔴	🔴	Text
PACO[11]	🟢	🟡	🟡	🟡	PBS
AWARE [151]	🟡	🟡	🟡	🟡	PBS
Ohmage [152]	🟡	🟡	🟡	🟡	PBS
Sensus [153]	🟢	🟡	🟡	🟡	PBS
PartS [101]	🟢	🔴	🔴	🟢	Visual
SurveySignal[12]	🔴	🔴	🔴	🔴	PBS
LifeData[13]	🔴	🔴	🔴	🔴	PBS
MovisensXS[14]	🔴	🔴	🔴	🔴	Visual
EthicaData[15]	🟡	🟡	🟡	🔴	PBS
Jeeves	🟢	🟢	🟢	🟢	Visual

🔴=Not implemented 🟡=Possible extension 🟢=Implemented

sub-question: *"What are the necessary features of apps that facilitate remote assessment of research participants or clinical patients in their natural environments?"* is partially addressed.

The 11 tools included in Table 4.4 are those which are currently available, with sufficient information to characterise them with regard to the four features of smartphone ESM derived in Section 3.2. Although the primary contribution of these tools is not necessarily their facilitation of ESM app development (for example, *AWARE* is intended as a tool for collecting smartphone sensor data) all support this functionality to differing extents. This section summarises the salient features and limitations of this state-of-the-art, with a brief description of each.

For the future reader, some of these tools may be in active development, such that the following descriptions, while accurate at the time of writing and information access, may now be obsolete.

4.2.2.1 iPromptU

iPromptU is an Android and iOS app that acts as both the software for developing an ESM app, and the app itself. The interface uses a forms-based (or PBS as described in Section 4.1)

[10]http://www.cogtherapy.com/iprompt u.htm Accessed 17/03/18
[11]http://www.pacoapp.com Accessed 17/03/18
[12]http://www.surveysignal.com Accessed 17/03/18
[13]http://www.lifedatacorp.com Accessed 17/03/18
[14]https://xs.movisens.com Accessed 17/03/18
[15]http://www.ethicadata.com Accessed 17/03/18

paradigm for specifying survey schedules. However, entering questions and schedules on the device itself could be cumbersome, particularly for studies with large banks of questions or complex triggering schedules.

A primary disadvantage of this approach is that running a study with multiple participants requires developing the study specification separately on each individual device. Despite running on modern smartphones with wireless capabilities, the data is stored on the app itself, so that devices must be returned to the researcher for upload and analysis.

4.2.2.2 PsychLog

PsychLog is a unique tool in this table, in its support for capturing physiological data as well as explicit self-reports, enabling researchers to make correlations between *explicit* emotional and *implicit* physical states. Physiological data is acquired through synchronising with an externally worn ECG monitor and tri-axial accelerometer.

PsychLog is designed for Windows Mobile which, at the present time, makes it inaccessible to the majority of smartphone users. Two disadvantages with regards to other state-of-the-art tools are its reliance on text file configuration of surveys, similar to *ESP*, and its inability to transfer data to researchers in real-time.

4.2.2.3 PACO

PACO is a web-based tool that allows development of ESM apps with various functions for both Android and iOS devices. Typical time-based sampling schedules are possible, as well as limited context-contingent sampling based on device usage, such as when a participant ends a call or opens another app.

The study creation interface of *PACO* displays all app functionality in a single web-based form which, while alleviating researchers from having to navigate multiple forms and pages, can quickly become cumbersome as specifications increase in detail. Nevertheless, *PACO* supports rapid development of basic, functional, cross-platform ESM studies.

4.2.2.4 AWARE

AWARE is presented as an Android framework for gathering, classifying and sharing contextual information. It has a wide range of data collection capabilities, with customised sampling frequencies and visualisations for all available smartphone sensors.

However, its ESM capabilities are a secondary feature and particularly limited, in that individual questions must be sent manually from the researcher's browser-based interface. Nevertheless, with its capacity to synchronise sensor data in real-time, and present meaningful visualisations to both researchers and participants, this combination of rich contextual and self-report data could be highly valuable for psychology research.

4.2.2.5 Ohmage

Ohmage is a platform allowing researchers to create participatory sensing campaigns, similar to tools discussed in Section 3.1.1. While participatory sensing apps are typically employed to collect sensor data on a large scale for answering particular research questions, *Ohmage* also supports sophisticated survey features for collecting self-report data from participants. Like *AWARE*, it has a variety of data visualisation features that can be accessed by both researchers and participants.

As a participatory sensing platform, however, a priority of *Ohmage* is maintaining participant engagement. Perhaps for this reason, it is not possible for researchers to define the conditions under which participants should receive survey notifications - each participant must do so on an individual basis. This means that, while time-contingent sampling is possible, this is at the discretion of the participant.

4.2.2.6 Sensus

Sensus is presented with a similar motivation to *AWARE*, as a platform for collection, aggregation and visualisation of participants' smartphone sensor data. The creation interface is a mobile app, which a researcher can use to configure a study's data collection parameters, self-report surveys, and the sampling schedule for these surveys. *Sensus* supports context-contingent sampling with almost all of the smartphone's sensors. However, as an opportunistic mobile crowd sensing application, rapid creation of ESM studies is not a primary focus.

The primary disadvantage of *Sensus* is the app-based interface for creating and modifying study specifications, similar to that of *iPromptU*. Moreover, with considerably more options to select from, the interface is particularly cumbersome to navigate.

4.2.2.7 PartS

Like *Ohmage* with regards to its motivation, *PartS* is designed and presented as a participatory sensing application. However, *PartS* appears to support a wider variety of ESM functionality, including context-contingent sampling as well as standard time-contingent schedules. The platform also integrates additional functionality, including communication between participants and researchers, a shared space to access visual feedback on participation and, of particular relevance to the work in this thesis, a means to construct context-contingent sampling schedules through a diagrammatic visual programming environment.

The publication describing *PartS* provides a wealth of information on its design process and useful features. However, as it is not currently available for access, it is difficult to identify disadvantages of the platform in terms of its usability.

4.2.2.8 Commercial Tools

While these tools have been developed and made openly available for use by researchers with budget constraints, the remainder of this section details proprietary versions that charge a fee for their use in studies. Such tools appear to provide a rich feature set and guided technical support. However, with no code base available, they are not open to modifications by researchers, thus in utilising these tools, researchers are still dependent on an external company for realisation of their study application. Moreover, replicating a study with a different set of participants, or sharing and collaborating with other researchers, are constrained by the "black-box" implementation of these systems, such that meta-design activities are discouraged.

4.2.2.9 SurveySignal

While the majority of modern tools utilise a native smartphone app, *SurveySignal* is a commercial tool that delivers assessments via SMS to participants. Researchers access the online interface of *SurveySignal* to tailor the content and delivery schedule of their survey prompts. The surveys themselves are delivered to participants as web links. However, researchers must create these surveys in an external tool, such as *Qualtrics*, and then use *SurveySignal* to send participants the relevant links.

As the only known example of this delivery mechanism in modern tools, the primary benefit of *SurveySignal* is its device agnosticism. Indeed, the only pre-requisite for participation is a mobile device with web connectivity. However, this flexibility also limits the functionality of the apps that can be created - no native smartphone app features, such as sensor usage, or capacity for self-initiated assessments, can be taken advantage of, for example.

4.2.2.10 LifeData

LifeData is another commercial tool that supports researchers in creating and deploying their own ESM studies to run natively on participants' Android and iOS devices. Like the majority of similar tools, both research and proprietary, it offers a simple web forms-based creation interface, wherein researchers can compose surveys, and define the sampling schedule.

LifeData has been designed to facilitate large-scale user research, and thus offers a scalable solution at varying costs. As with other PBS tools, its functionality appears to have been simplified in order to maximise usability for non-programmer researchers. As such, it does not offer any of the features suggested in Section 3.2. For example, only GPS sensing can be employed with no means to specify under what circumstances, and cannot be used as a trigger condition for context-contingent sampling.

4.2.2.11 MovisensXS

A particularly interesting feature of *MovisensXS* is its diagrammatic visual programming paradigm, the only commercial tool that uses such an approach. *MovisensXS* allows researchers to create study specifications by dragging boxes that represent constructs such as assessments and sampling schedules onto a canvas, and connect them together with lines, similar to the paradigm employed in *PartS*. The resulting study specification resembles a flowchart that visually represents the functionality of the application.

Although no statistics are available, *MovisensXS* appears to be the most popular state-of-the-art platform in research, with its use cited in seven of the recent studies surveyed in Chapter 2. However, its flowchart paradigm does not appear to have been evaluated in any usability studies, such that it is currently unclear whether this is an intuitive paradigm, or whether significant development support is required for research to realise semantically correct study apps.

4.2.2.12 EthicaData

EthicaData is a commercial platform that allows researcher developers to implement their own study-specific features in Java through the *Ethica* API. It has further features of potential benefit to researchers, including the capability to capture various types of sensor data, which can then be correlated with explicit self-report.

Of the commercial tools, *EthicaData* appears to support the widest range of functionality. However, modifying individual surveys and triggers requires researchers to navigate a hierarchy of menus and forms within the web-based editor, such that it could be difficult to obtain a summary of any given study, and making small changes to a large study could be particularly time-consuming.

4.2.2.13 Summary

Existing tools in both research and commercial domains each have their own individual advantages and disadvantages. A particular limitation common to these tools is their lack of flexibility in allowing researchers to capitalise on the features derived in Section 3.2, which will be discussed in the following section. Although usability issues were encountered in personal use of the systems, no formal usability evaluation of these systems was undertaken, thus these issues are anecdotal and not evidence-based criticisms.

While the commercial tools offer features beyond those available in Jeeves, a particular disadvantage of these tools is that researchers have no ownership of their studies, and are thus inhibited from sharing and reuse of their created specifications. In the ESM studies reviewed in Chapter 2, a common concern was that studies were limited to a certain group of participants, or that potential variables were omitted from the data collection procedure, or other such limitations that would merit running the study again.

4.2.3 Features and limitations

As a result of this review, the need arises for an intuitive, empowering solution that allows researchers to create, deploy and collect rich data from their own ESM studies from beginning to end, with full flexibility to adjust their studies without the constraints of a commercial system. Useful features identified in the previous chapter were chosen as comparison measures across the existing EUD-ESM tools. The presence, or absence, of the capacity for tools to support these features are discussed as follows.

4.2.3.1 Implicit Context Triggers

Tools that enable specification of context-contingent sampling schedules, identified as a particularly useful feature in both psychology and clinical research, present limited functionality. While the commercial tools *LifeData*, *MovisensXS* and *EthicaData* enable the GPS tagging of self-reports, none of these tools appear to enable researchers to trigger based on this location, or indeed other sensors. *EthicaData* does provide an API through which developers can create and link their own trigger functionality, but doing so requires Java programming experience. As open-source software, *AWARE* and *Ohmage* could be programmed to enable sensor-based triggering, but do not include this functionality in their available state.

Of the tools that allow context-contingent sampling, *Sensus* is perhaps the most diverse in its triggering capabilities, supporting external devices as well as a range of built-in sensors. The interface for defining trigger conditions within *PartS* additionally provides user interfaces that

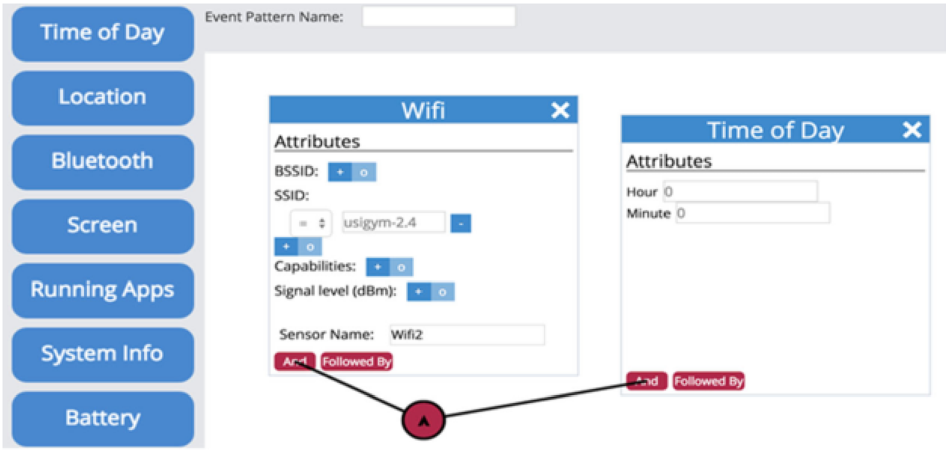


Figure 4.5: The “event pattern editor” of *PartS*, allowing combined trigger conditions

enable combinations of trigger conditions, as shown in Figure 4.5. In this example, a trigger is being defined that fires when a WiFi network is detected at a particular time of day.

4.2.3.2 Explicit Context Triggering

None of the tools surveyed allow for **explicit context triggering**, that is, performing actions that are contingent on attributes of a participant, derived from responses to surveys, for example. Such functionality would be necessary in order to provide tailored intervention feedback to participants. However, self-report data in all reviewed tools cannot be interpreted by the apps themselves, such that any intervention would need to be triggered by the researcher manually after reviewing participants’ data.

4.2.3.3 Preference Tailoring

Similarly, as evidenced by Table 4.4, functionality to automatically tailor ESM to the preferences and characteristics of individuals is not currently supported in EUD tools. In using *ESP*, one of the first examples of electronic ESM, study-specific PDA devices were manually programmed to account for participants’ waking and sleeping times. Although tailoring to participants’ schedules can improve compliance [103, 104], manually doing so is a burdensome process for researchers, and inflexible to changes in participants’ schedules. *Ohmage* and *PartS* allow participants to set their own reminders, but researchers have no flexibility to add other customisations.

4.2.3.4 Two-Way Feedback

Communication functionality is understandably limited in existing tools. Given the requirements of anonymity and consistency inherent in ESM research studies, direct researcher-participant communication has potential ethical implications, and biased communication delivered to some participants, but not others, could bring the validity of collected data into question. Some tools enable study information messages to be sent to all participants, but with no way for these participants to provide feedback, or to have individual interactions.

Nevertheless, as a participatory sensing platform, *PartS* supports two-way communication as a motivational incentive for participants. Such functionality is supported by feedback from an end-user evaluation by the platform's authors, which showed that participatory sensing researchers had *"a significant degree of interest in maintaining contact with other participants or researchers"* [101, p. 496]. This suggests that, while an uncommon feature of experience sampling studies, its applicability as a means to increase compliance and obtain feedback on a study during its course could make it highly desirable in certain contexts of use.

4.2.3.5 Creation Interface

The interface for configuring ESM specifications most frequently comprises web forms, which abstract over a semi-structured language description. As discussed in Section 4.1, this is defined as "programming by specification" (PBS). Such an interface provides a more approachable solution than the pioneering platforms mentioned previously, wherein researchers modify study specifications through plain-text or XML files.

Two notable exceptions are *MovisensXS* and *PartS*, which both utilise a visual programming environment that abstracts study concepts into graphical components. A visual approach can support end-users in configuring trigger-action behaviours, as demonstrated by Dey et al. in the development of the *iCAP* application [154]. Both *MovisensXS* and *PartS* employ a diagrammatic visual language for editing studies. While *PartS* allows for simple triggers to be specified without linking the individual "event" boxes illustrated in Figure 4.5, *MovisensXS* requires that the entire study is specified as a flowchart.

It is notable that none of the tools employ "programming by demonstration" (PBD) to any degree. While PBD is an intuitive means of demonstrating context-aware behaviours, as shown in tools including *a CAPpella* [120], *GALLAG Strip* [137], and *Context Studio* [155], these are all examples of "private EUD" (development for personal use). In the development of apps for a range of participants, it is infeasible for a developer to demonstrate the personalised contexts of all participants in advance.

4.2.4 Summary

Many EUD-ESM tools, as well as bespoke ESM apps, were discovered in the review reported in this section, all of which offer their own unique contributions or features, but none of which support the full set of features derived in the literature review. Commercial tools offer a variety of polished features for data visualisation and presentation of survey questions to participants. However, without any published studies to determine what researchers would desire from EUD-ESM tools, it is difficult to determine how useful such features are to these researchers. Other platforms such as *AWARE* and *Sensus* focus on the collection and visualisation of contextual data. While highly efficacious in doing so, they are not specialised solutions for conducting ESM studies and as such face issues in usability and flexibility when employed as such.

In all cases the systems reviewed have a low barrier to entry, allowing simple studies to be set up with minimal effort. However, as shown in Table 4.4, flexibility is sacrificed, with the inability for researchers to utilise a number of potentially useful features. In particular, the ability for researchers to provide feedback based on previous survey responses, or to tailor to individual participants, is not present in any of the reviewed systems. Although usability is clearly a factor in the development of these tools, only two (*Sensus* and *PartS*) report findings of a conducted usability evaluation. In terms of real-world utility, only the *Ohmage* publication reports on field deployments where the tool has been used in practice.

4.3 Design Implications

While the previous chapter derived a set of design implications for smartphone ESM tools, the implications of this chapter encompass EUD tools more broadly. They are not specific, prescriptive guidelines for such tools, but instead represent decisions that were taken for the implementation of Jeeves, based on the reviewed literature. These decisions are summarised as follows.

Although ensuring a “gentle slope of complexity” is a key design principle towards education-based EUD environments, **as there are discernible groups of potential software modifiers, a Software Shaping Workshop (SSW) approach should be employed in Jeeves**. This ensures that only development activities of relevance are permitted for each group of users.

While the SSW approach minimises learning effort through only allowing relevant actions to be taken, the programming paradigm is also critical to EUD success. Blocks-based visual programming has been shown in various workshops with children and adolescents to be intuitive and accessible, while supporting a high degree of flexibility. Thus, **a blocks-**

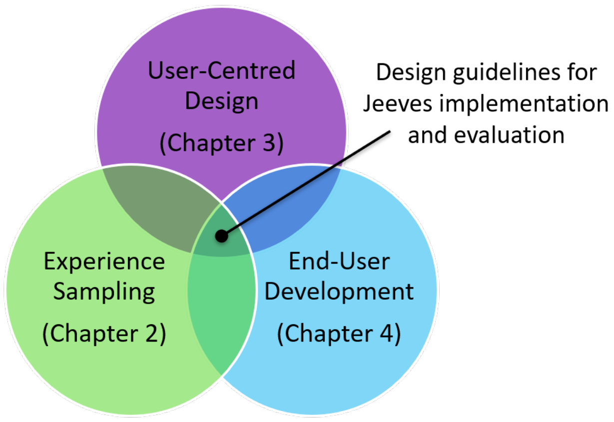


Figure 4.6: The design implications of Chapters 2, 3 and 4 guided the work of successive chapters

based programming paradigm should be employed in Jeeves to provide flexibility to non-programmer developers.

Evaluation of EUD in its intended real-world context is limited, with the majority of evaluations taking the form of lab-based usability studies. Irrespective of the design of the EUD tool itself, **subsequent user evaluations should be designed to ensure that Jeeves matches its intended usage context**.

No existing EUD tool for ESM app creation implements all four design features derived in the previous chapter as design implications. Moreover, the features implemented are seldom evaluated. Thus, in order to evaluate these features for their perceived utility, the final design implication arising from this chapter is that **the four features derived from Chapter 3 should be implemented in Jeeves**.

4.4 Chapter Summary

The background work reviewed in this chapter introduces end-user development as a core concept of this interdisciplinary research. In particular, the final section described developments where this has been united with the other core concept of experience sampling, highlighting limitations that should be addressed in future EUD tools for creating ESM apps, which are presented in the form of design implications. From the literature reviewed and analysed in this chapter, as well as that in Chapters 2 and 3, the remainder of this thesis involves the design, implementation, and evaluation of a blocks-based programming environment for EUD of ESM apps. As illustrated in Figure 4.6, design implications from these three chapters have informed the design of this environment, named Jeeves, which constitutes the tool with which the overall research question can be addressed - identifying the factors influencing the adoption of EUD for ESM apps.

CHAPTER FIVE

DESIGN OF JEEVES

Ideas in this chapter have been published in the following papers:

D. Rough and A. Quigley, "Overcoming mental blocks: a blocks-based approach to experience sampling studies," in *Blocks and Beyond Workshop (B&B), 2017 IEEE*, pp. 45–48, IEEE, 2017.

D. Rough and A. Quigley, "TAPping into mental models with blocks," in *Blocks and Beyond Workshop (B&B), 2017 IEEE*, pp. 115–116, IEEE, 2017.

The previous chapter described the concepts of end-user development (EUD), and their relation to the development of ESM applications. In particular, it outlined how the current approaches to EUD of ESM applications are, in many ways, insufficient for addressing the challenges of these applications' domains. With respect to the limitations of existing EUD tools, this chapter presents Jeeves, a blocks-based visual programming environment, developed as a prototype EUD tool to fulfill the requirements outlined.

This chapter begins with a brief overview of the Jeeves interface, providing a basic description of the various components. Following this, from the related work on previous and existing EUD-ESM tools and recent ESM studies, this chapter formalises a set of design requirements that represent the baseline functionality for a useful EUD-ESM tool, as well as those for an EUD-ESM tool that improves upon state-of-the-art platforms. The design of Jeeves is described in greater detail in relation to each requirement, including novel concepts of *meta-tailoring* and *event-state interventions* that have been implemented as a result of these requirements. The visual programming language itself is further described, with design features derived from two sets of guidelines for visual notations. Finally, the design of Jeeves as a full software platform is

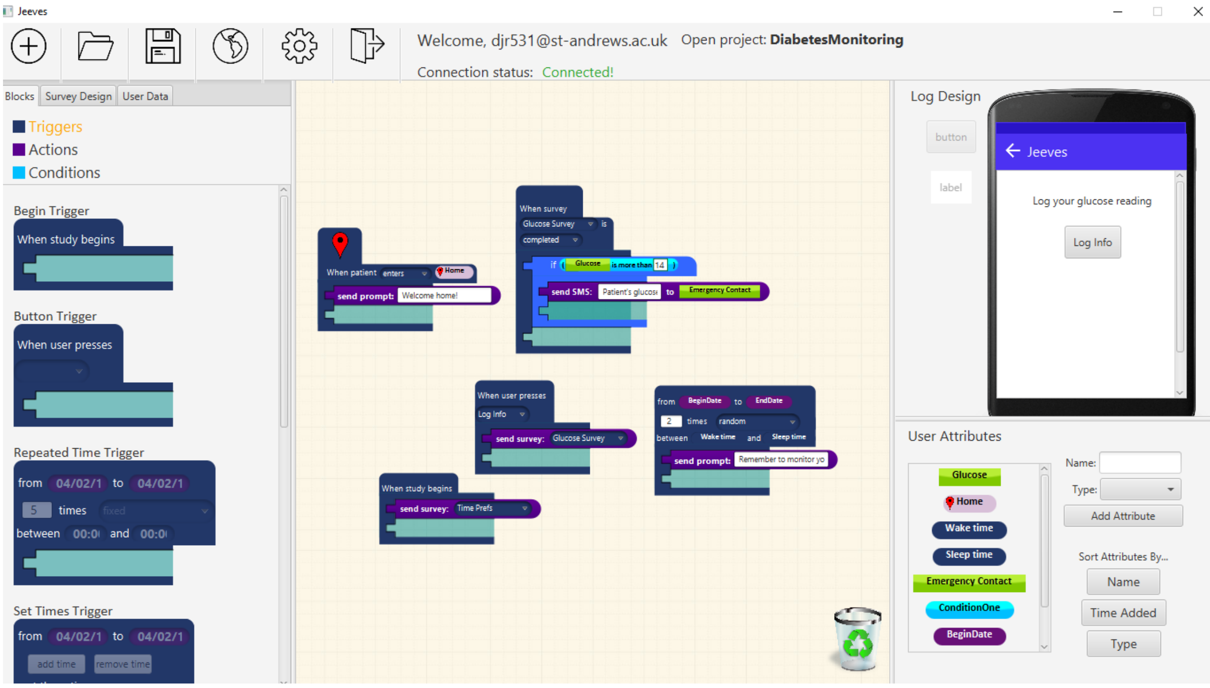


Figure 5.1: Main view of the Jeeves EUD tool

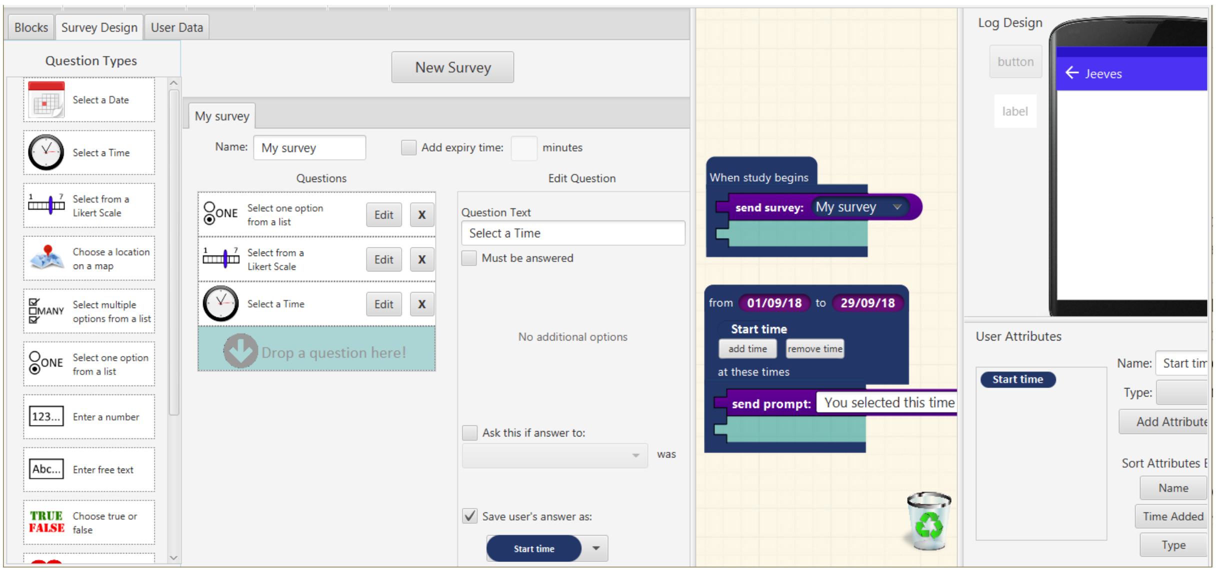


Figure 5.2: Survey design view of the Jeeves EUD tool

described, including the JeevesAndroid app, and server-side functionality that links this app with the visual programming environment.

5.1 Introduction to Jeeves

This section provides a brief overview of how a researcher or clinician (hereby referred to as a developer) can interact with the Jeeves EUD tool, with reference to Figures 5.1 and 5.2. Primarily, the *Blocks* view of Jeeves is shown in Figure 5.1. The centre of the image shows the canvas onto which the visual blocks can be dragged, connected together, and rearranged. These blocks are organised into different menus by the function which they provide, and are displayed on the left side of the screen. In this example, the *Triggers* menu is shown, from which developers can drag and drop trigger blocks (identified by their dark blue colour) of choice to structure their application. Action blocks and Condition blocks are also contained within their own menus, which a user can toggle between using the selector in the top-left of the screen.

The top-right of the screen shows the *Log Design* pane, where a user can drag and drop buttons and labels onto a simulated smartphone screen. In doing so, the developer can create and describe buttons that perform certain actions when pressed. In the example shown, the developer has created a “Log info” button, which will appear on the end-user’s app screen.

Finally, the lower-right section of the screen shows the *User Attributes* panel, from which a developer can create new blocks that represent variables and individual characteristics of a participant. By creating and incorporating these Attribute blocks into the existing block specification, functionality specific to each end-user can be implemented. The basic principle is that the values of these attributes are set by participants through their answers to in-app survey questions, thus permitting the logic of the application to be unique to each participant.

Figure 5.2 shows the *Survey Design* section of Jeeves, through which a developer creates and modifies the surveys that a participant interacts with. This simple example also demonstrates how surveys are related to blocks and attributes. Similar to the Blocks view, question blocks are dragged from the *Question Types* menu on the left of the screen into the *Questions* list. Parameters of a selected question can then be modified in the *Edit Question* subsection. In the example, the “Select a Time” question has been modified so that a participant’s answer to this question is saved into the “Start time” attribute. Additionally, the small section of the canvas that can be observed shows how a developer can instruct this survey to be sent by using the *send survey* action with the particular survey’s title. The lower trigger also shows how the time-based attribute is used to send a prompt at the participant’s chosen time, rather than a static time defined by the developer.

Screenshots and example blocks of previous versions of Jeeves are illustrated in Appendix D.

5.2 Design Requirements

Designing a successful EUD tool is a process that requires balancing the functionality of the tool with its ease-of-use by non-programmers. A complex tool that provides maximum flexibility for ESM creation would likely raise the same issues encountered by researchers being forced to learn a programming language from first principles. Conversely, a tool with a sole focus on ease-of-use will not allow researchers to realise the more complex functionality that constitutes a large contribution of this work. Design guidelines for EUD-ESM tools have been proposed in previous work: for example, Fischer derived guidelines from a review of three tools that, based on the technology of 2009 in which the guidelines were published, are now largely obsolete [156].

This section will detail a set of design requirements synthesised from related work, and explain how these have been implemented in Jeeves. This is not a comprehensive requirements analysis from a software engineering perspective; rather, their purpose is to structure the description of Jeeves, and to guide implementations of similar tools. Moreover, the requirements of each section are divided into requirements that *must* be implemented as a minimum standard in every EUD-ESM tool, and requirements that *should* be implemented (those that Jeeves satisfies that existing EUD-ESM tools do not). Foremost, the sampling strategies that developers can employ are expressed in terms of design requirements as follows.

5.2.1 Sampling strategies

The sampling strategies available in an EUD-ESM tool guide the timing and frequency of assessment requests. The trigger hierarchy illustrated in Figure 5.3 categorises the events upon which actions can be taken during an ESM study. These are broadly divided into time-based and event-based triggers, as described below. Definition of the conditions under which assessments are prompted, or other actions are taken, are key to the design of useful ESM studies.

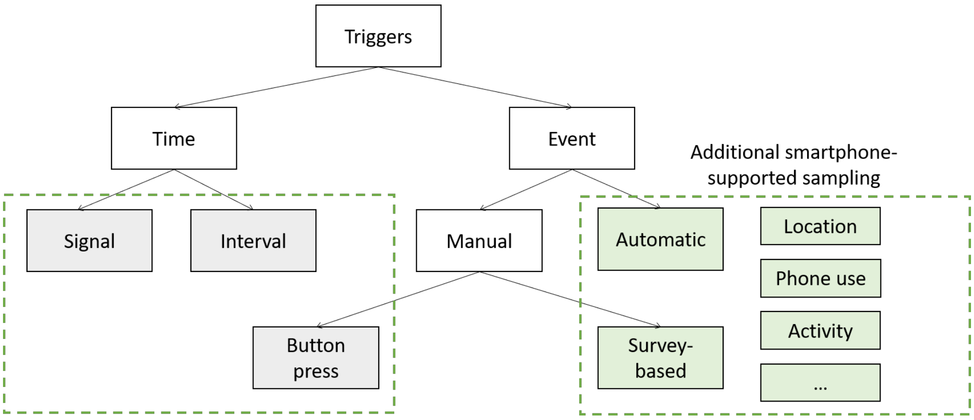


Figure 5.3: A taxonomic illustration of triggers incorporated in Jeeves

Table 5.1: Trigger blocks enabling minimum required sampling

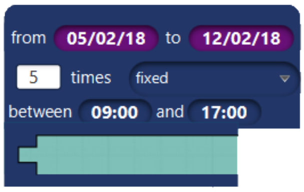
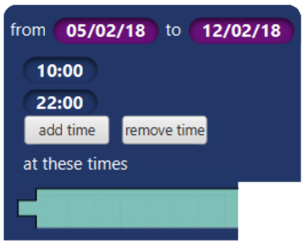
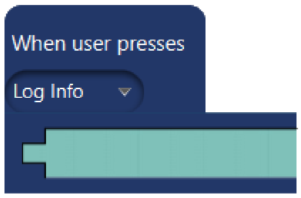
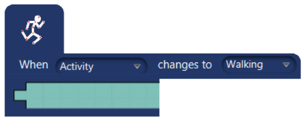
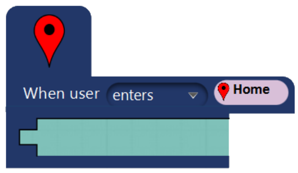
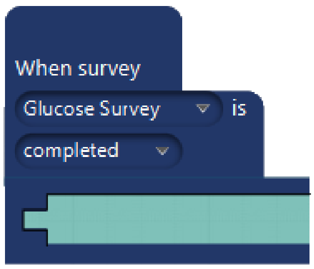
Sampling Strategy	Jeeves Representation
Signal contingent (pseudo-random times)	
Interval contingent (discrete times of interest)	
Event contingent (user prompted events)	

Table 5.2: Trigger blocks enabling further context-aware sampling

Sampling Strategy	Jeeves Representation
Sensor-based (Activity, phone usage, etc)	
Location-based (User enters/ leaves a place)	
Survey-based (User completes/ misses a survey)	

Requirement 1 - Jeeves must support signal, interval, and event-contingent sampling

The sampling strategies proposed by Wheeler and Reis, as described in Chapter 2, represent the minimum functionality of an ESM study [47]. In Jeeves, trigger blocks exist for each of these three sampling strategies, allowing researchers to specify a time window and frequency of random sampling (**signal-contingent**), or specify discrete times of interest (**interval-contingent**).

Further, user-initiated **event-contingent** sampling is possible through triggers based on user interface button interactions. The relevant blocks are shown in Table 5.1.

In signal-contingent sampling, randomness is generally controlled by ensuring that signals occurred at least once within each time window, often with a minimum time distance apart. This 'pseudo-random' sampling ensures that participants cannot anticipate an alert in advance, while also ensuring that clustered alerts do not occur.

Requirement 2 - Jeeves should support context-sensitive sampling

As indicated in the taxonomy shown in Figure 5.3, two additional sampling strategies have been added that were previously impossible with the paper/signalling device combination assumed by Wheeler and Reis. These are *sensor-based triggers*, and *survey-based triggers*. The blocks for these new types are shown in Table 5.2. These triggers are direct implementations of two design implications established in Chapter 3, namely that **implicit data collection from smartphone sensors should be incorporated** and **prompts based on participants' unique responses should be incorporated** respectively.

Sensor-based triggers are those that exploit the integrated sensors of a smartphone. Recent ESM studies have employed the acquisition of location, activity, or phone usage for triggering surveys or prompts. However, with the growing array of sensors integrated into today's smartphones, new automatic sensor triggers could be added if a use case arises. For example, recent models of smartphones can sense heart rate, and the implementation of an ESM creation tool should allow new sensor classifiers to be easily integrated by developers. While the review in Chapter 2 shows that the uptake of context-sensitive sampling in psychology and medicine is low, van Berkel et al. highlight its prevalence in computer science research [56]. The utility of context-sensitive functionality was first recognised by Intille [57] as having major benefits to psychology research, thus up-to-date EUD-ESM tools should allow researchers to implement this functionality with minimum effort. *Geofencing* triggers, those contingent on location, appear particularly prevalent in literature and as such has been represented as a separate block with more fine-grained options than standard sensor-based triggers. For example, triggers can fire based on a user entering, leaving, or dwelling in a particular location.

Despite the wealth of sensors available on modern smartphones, little has been done to allow researchers to define their own context-aware ESM apps. This is most likely due to three factors - the difficulty of classifying raw sensor data, restrictions on sensor access and background processes in smartphone operating systems, and, perhaps most importantly, the drain of continuous sensing on battery life. However, resource-intensive sensing is not necessary to enable useful context-aware functions. For example, Ho and Intille found that a person's interruptibility was increased when they were transitioning from one activity to another, which was detected by a simple change in acceleration [84]. A more recent example is the *InterruptMe* library by Pejovic and Musolesi that learns interruptibility based on a dataset of sensed and self-reported ESM data [157].

While not reliant on *implicit* context, survey-based triggers react to a participant completing a survey, or failing to complete a survey within its allotted time. As such, it is considered to be a context-aware trigger, in that it is contingent on the *explicit* context of participants. Such functionality does not impinge on battery life, as no complex classification algorithms are required. However, it still has potentially useful applications in ESM apps. For example, compliance prompts can be issued when a survey is missed, or intervention feedback delivered automatically if a participant has completed a survey with particular answers, as described by Requirement 7 in Section 5.2.3.

5.2.2 End-user tailoring

Requirement 3 - Jeeves must support tailoring to individual participants

Another important requirement not addressed by existing EUD-ESM tools is that of tailoring to individual participants, which has influenced many decisions on the design of Jeeves. In the review of recent ESM studies in Chapter 2, many have sampling schedules that are individually tailored to participants' waking and sleeping times. User-centred design studies reviewed in Chatper 3 also advocate the ability to personalise reminders based on participants' desired times or locations. Consequently, the design implication that **a means for participants to tailor apps to their personal preferences should be incorporated** was established in this chapter. *Ohmage* [152] and *PartS* [101], two of the alternative EUD-ESM tools discussed in Chapter 4, do this to some degree by allowing participants to set their own reminders at personal times and places to ensure their own compliance.

Tailoring to individual participants should improve compliance, given that a fixed sampling schedule may not fit with every participant's waking hours. Indeed, in implementations using *ESP* [46], the first EUD-ESM tool, personal schedules were programmed separately for each participant - a tedious process when many participants are involved, and inflexible if any participant wishes to change their schedule. Another use case for storing personal information could be to detect a *semantic* context, such as a participant's home address, as opposed to a more generic change in GPS location. Since home and other semantic locations are specific to each participant, an EUD-ESM tool must have some level of tailoring capability, such as that implemented in *Ohmage* and *PartS*. However, with this approach, an interesting tension arises,

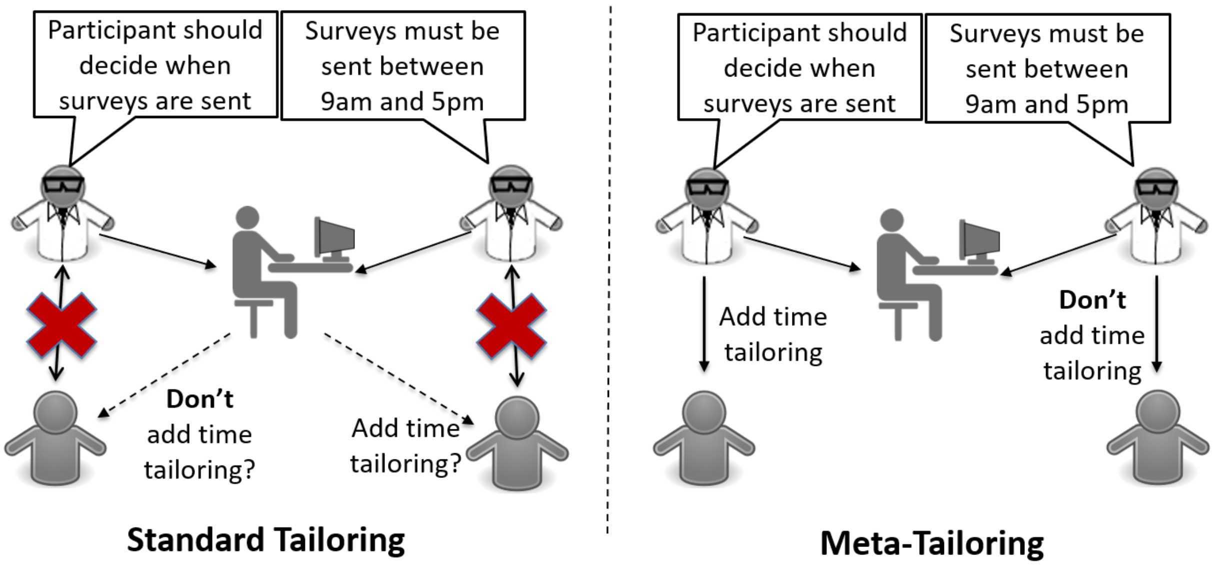


Figure 5.4: Illustration of the difference between standard tailoring and "meta-tailoring"

in that the needs of the researcher may be compromised by the needs of the participant if too much flexibility is afforded, discussed as follows.

Requirement 4 - Jeeves should allow researchers to define the level of tailoring

The underlying justification for meta-design is that software developers cannot anticipate the variety of requirements that end-users will have, therefore these users should become developers of their own tailored software [139]. However, it also becomes apparent that researchers and clinicians, acting as software developers, cannot anticipate the constraints of their participants and patients. Compliance issues result from prompts that come at inconvenient times and places. Moreover, geofencing triggers in EUD-ESM tools such as *MovisensXS* require a specific set of GPS coordinates, and fail to account for conceptual locations individual to a participant. For example, what is 'home' to one participant is unlikely to be the same for another. Thus, researchers should be able to add functionality that allows participants to tailor these attributes themselves. This is the definition of **meta-tailoring**, and is illustrated in Figure 5.4.

When end-user developers are not the actual end-users of the application itself, this is a *"multi-tiered proxy design problem"* as defined by Fogli and Piccino [158]. In modelling this problem, they consider the developer acting as a proxy for the end-user, who is incapable of designing the software structure, but may be allowed flexibility through setting particular parameters. They argue that, *"thanks to meta-design and meta-modeling, such systems may have some flexibility degrees, and thus be personalized by end users"* [158, p. 165]. This view of end-users, while correct, is limited to suggesting that tailoring functionality should be decided at the *meta-design*

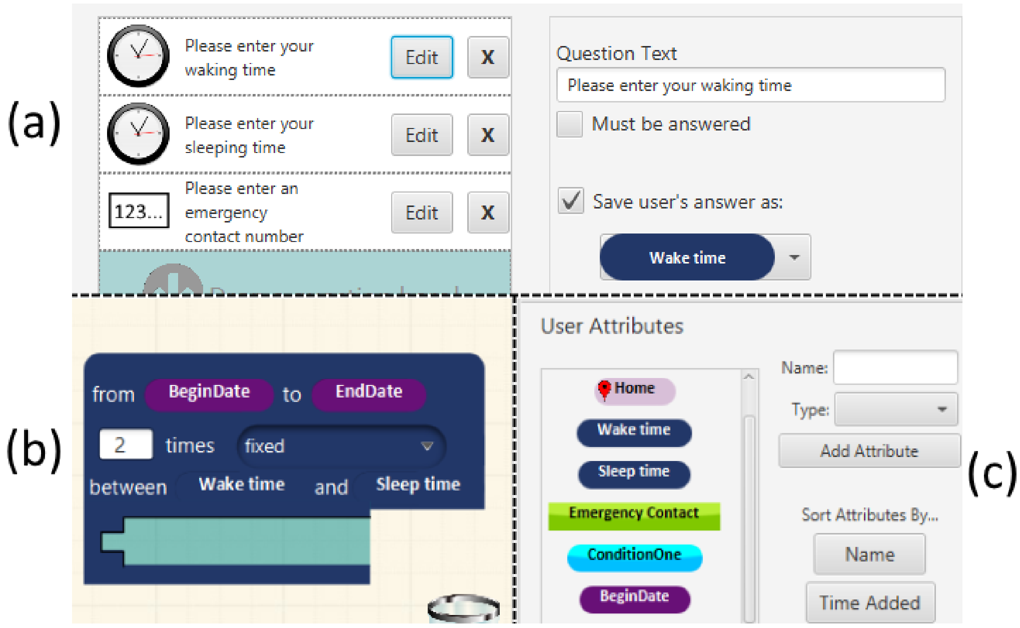


Figure 5.5: Researcher designs a survey question, used to set an attribute value that tailors a trigger

phase. Instead, meta-tailoring has this decision made at the *design* phase by the researcher.

Consider Figure 5.4. One researcher requires an app where participants can tailor their survey time window. Another researcher requires a more rigorous design where surveys are sent specifically between 9am and 5pm. Should the meta-designer add a time tailoring feature? The left scenario represents the current multi-tiered proxy design problem - whether the meta-designer adds this feature or not, this conflicts with the needs of one of these researchers. On the right represents the meta-tailoring concept - the researcher can add this feature or not, dependent on their study requirements.

User attributes provide the individual tailoring features of Jeeves. Rather than having to manually configure their preferences, this can be done automatically. The *User Attributes* pane shown in Figure 5.5(c) can be used to specify personal attributes of a participant, acting as programmatic variables. Attributes can be numbers, true/false values, times, dates, and locations, which can then be used to tailor study configurations. For example, in Figure 5.5(b) the trigger is tailored to the individual dates and times of the participant. The participant can set the values of these attributes through providing answers to survey questions. As shown in Figure 5.5(a), a participant's answer to a question can be saved into an attribute. In this example, a researcher can create a survey to obtain participants' waking and sleeping times, that store values specific to each participant. Attributes, the question types they can be assigned from, and expressions with which they fit, are illustrated in Table 5.3. (Expressions are further explained in Requirement 7.) In summary, by incorporating these two levels of tailoring at the participant and researcher level

Table 5.3: Attribute types, survey questions they can be assigned from, and relevant expressions

Type	Appearance	Assigning Questions	Compatible Expressions
Boolean	ConditionOne	TRUE FALSE Choose true or false	(ConditionOne and) (or ConditionOne) (ConditionOne is false)
Numeric	Glucose	123... Enter a number 1 7 Select from a Likert Scale	(Glucose is equal to) (is more than Glucose) (Glucose is less than)
Category	Mood	ONE Select one option from a list MANY Select multiple options from a list	Mood is equal to Happy
Date	BeginDate	Select a Date	BeginDate to 26/02/18 1 month before BeginDate
Time	Wake time	Select a Time	Wake time to 22:00
Location	Home	Choose a location on a map	user is at Home

respectively, this addresses Chapter 4's design implication that **a Software Shaping Workshop (SSW) approach should be employed in Jeeves** .

5.2.3 Assessment and intervention

The requirements of assessment are generally straightforward, encompassing a small number of question types, with conditional branching logic to skip over unnecessary questions. However, *interventions* and appropriate ways to trigger them are more difficult to define.

Separate from the blocks specification, but still an integral part of Jeeves, is the Survey Creation Pane, shown in Figure 5.6. This drag-and-drop interface was designed from a review of existing EUD-ESM tools, as well as online tools for survey creation such as *SurveyMonkey* and *Qualtrics*. Different question types can be customised with a simple form. The current implementation supports conditional branching on previous answers, and the ability to store a participant's question response as the value of an attribute, as shown in Figure 5.5. The reason for such

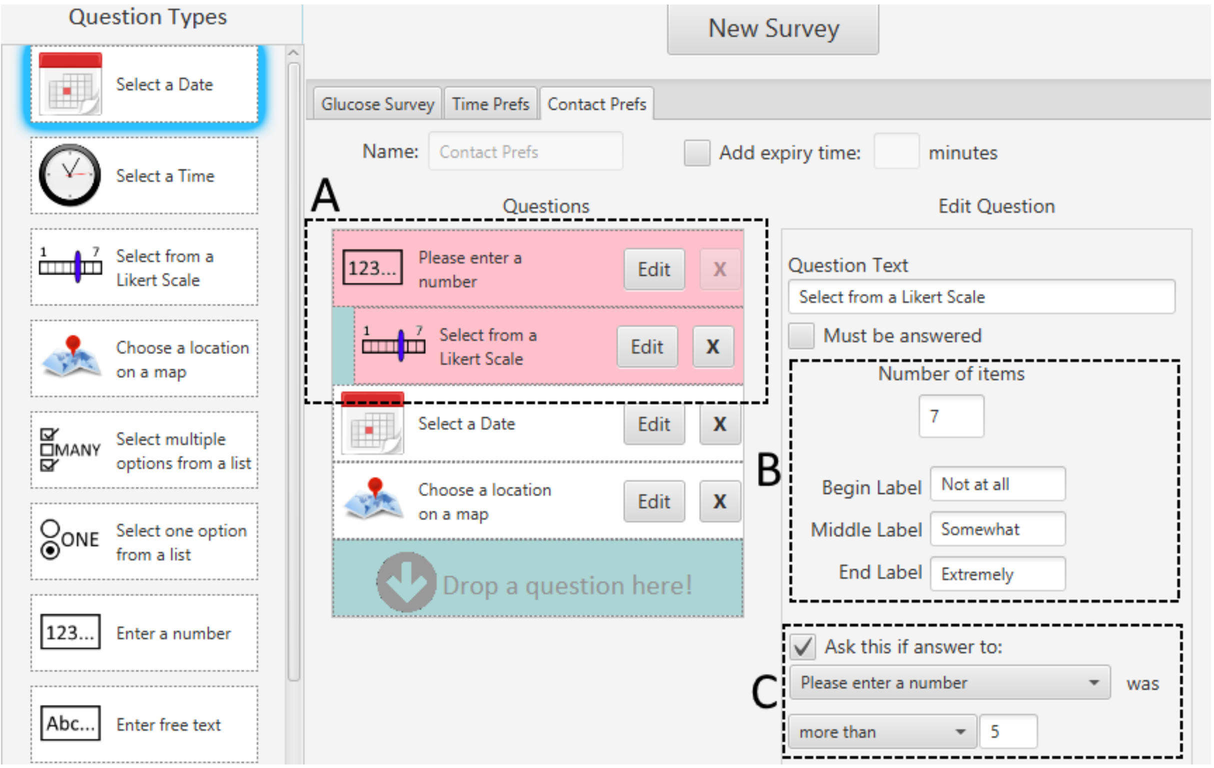


Figure 5.6: A screenshot of the Jeeves Survey Design pane

features are explained in the following requirements.

Requirement 5 - Jeeves must support a variety of survey question types

The types of question asked in psychology and clinical research are free-text questions, single selection questions, multi-selection questions, and Likert scale questions. Variations of such questions are also necessary (for example, questions restricted to numbers, or true/false questions). Jeeves provides additional question types that store dates, times, locations, and heart rate.

Requirement 6 - Jeeves must support conditional branching based on question responses

To minimise the burden on participants, it is necessary for EUD-ESM tools to implement some form of branching that jumps over irrelevant questions according to a participant's response to an earlier question. In Figure 5.6, the left side shows the variety of available question types, with the middle showing questions currently integrated in the selected survey. The darker shade of the top two questions (Box A) shows that these questions are related. The indentation of the lower question denotes that it is conditional on the answer to its predecessor.

Certain question types require additional customisation (Box B). For example, Likert scale questions can have a variety of labels, and the variance in the scale differs considerably between studies. The conditional branching function can be seen (Box C), whereby certain questions with predictable answers within a range (such as those with numeric responses, or responses from a set list of options) can act as branching questions that determine whether future questions are asked. In the Figure 5.6 example, the Likert scale question is only asked if the response to the numeric question is more than 5.

Requirement 7 - Jeeves should support tailored interventions

The requirement for intervention capabilities in research and clinical practice is particularly salient, yet is not addressed in current EUD-ESM tools. In Heron and Smyth's seminal paper on ecological momentary interventions (EMIs) they describe intervention feedback that could be delivered on a number of conditions [75]:

1. On completion of an experience sampling survey
2. Prompted by a participant in response to an event
3. Triggered in response to external context

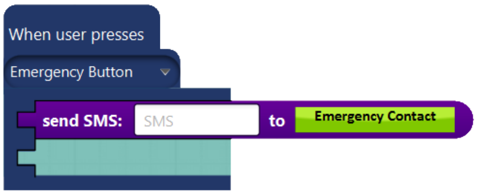
The intervention content, in both psychology research and clinical practice, typically consists of a supportive message or provision of information. While less appropriate in controlled research studies, practice-based interventions can also solicit the involvement of an external individual, such as a carer or a healthcare provider. The ability for an app to communicate with a chosen contact is a desirable feature for particular contexts (such as apps for monitoring heart failure or other life-threatening conditions [89]). Thus, information prompts should be deliverable to both a study participant, and to suitable contacts with a stake in the participant's well-being.

Heron and Smyth further describe how mobile implementations can "*allow for the timing of messages to be individually tailored by providing EMI at specific times when participants or patients are most in need of additional support*" [75, p. 3]. In doing so, participants are more likely to respond favourably to an intervention, and seek this support with minimal delay.

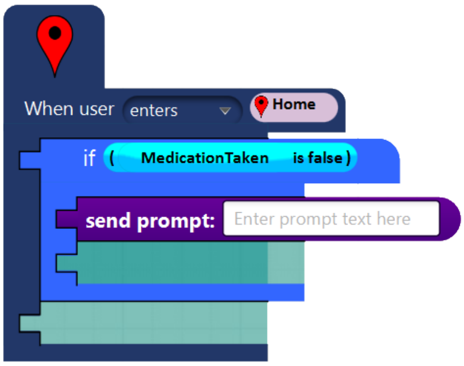
Table 5.4 describes three Jeeves implementations of example interventions, which include a number of different blocks beyond the basic introduced triggers and actions. While the trigger blocks are used to represent discrete events, the if-condition and expression blocks are used to represent continuous, ongoing state - a conceptual difference explained by Ur et al. [159]. Thus, Jeeves supports tailored intervention delivery through **event-state interventions**. The lower two examples demonstrate this combination of event and state. The discrete, one-time *event* (such as a participant entering a new location or completing a survey) does not alone warrant execution of intervention actions. However, if the participant is in a particular *state* at the time of the event, represented by the **if-condition** and its conditional **expression**, then the intervention actions

Table 5.4: Types of intervention and their Jeeves representation

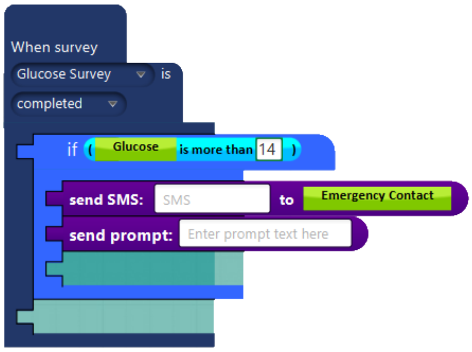
<table>
<tr><th>Sampling Strategy</th><th>Jeeves Representation</th></tr>
<tr><td>User presses a ‘panic button’, SMS is sent to an emergency contact</td><td>



</td></tr>
<tr><td>User arrives home. If they have not taken their medication, a reminder prompt should be sent</td><td>



</td></tr>
<tr><td>User completes glucose assessment. If their reading is above a defined value, a message is sent to the participant and emergency contact</td><td>



</td></tr>
</table>

should be executed. These two block types are further described.

5.2.3.1 If-Conditions

In combination with triggers, if-condition blocks specify additional constraints under which to execute an action. The notch on the if-condition block shows that it acts as a form of action, and can be logically nested within triggers. Unlike other actions, if-conditions have their own receiver for further action blocks that should be conditionally executed. This also means that if-condition blocks can be nested within one another indefinitely, an example of which is shown in the lower-left of Figure 5.7. To determine whether the nested actions of an if-condition are executed, the constraint is represented by an expression block.

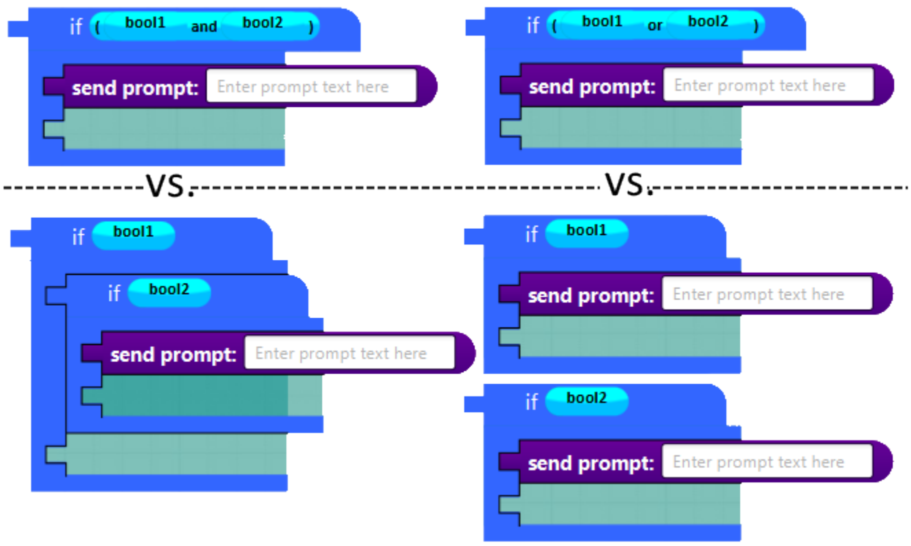


Figure 5.7: Expression blocks reduce the verbosity of the blocks paradigm

5.2.3.2 Expressions

Expression blocks are used to translate various user attributes into Boolean expressions. (The expressions available for each attribute type are shown in Table 5.3.) Additionally, they can function to simplify complex conditions, thereby reducing the number of if-condition blocks required. For example, Figure 5.7 shows how the *and* expression and the *or* expression in the top two conditions can be used to represent the multiple conditions below it.

5.2.4 Communication

Communication between participant and researcher, or more commonly, between patient and clinician, can facilitate compliance, ensure health problems are resolved promptly, or provide reassurance and support to distressed participants. However, given the constraints of anonymity and consistency in research studies, communication between researcher and participant in existing EUD-ESM tools is not commonly addressed.

Requirement 8 - Jeeves must let researchers monitor participant data in real-time

Smartphones enable data from surveys to be transferred to researchers instantly. Indeed, almost all the surveyed tools in Chapter 4 allow researchers to view incoming data in real time, and make modifications to a study as necessary. For enabling the real-time monitoring of participants, and modification of study protocols, the Jeeves architecture uses a basic client-server model, with the *Firebase* online database at the back-end, and the desktop application and Android app acting as the researcher and participant clients respectively. This architecture is described in

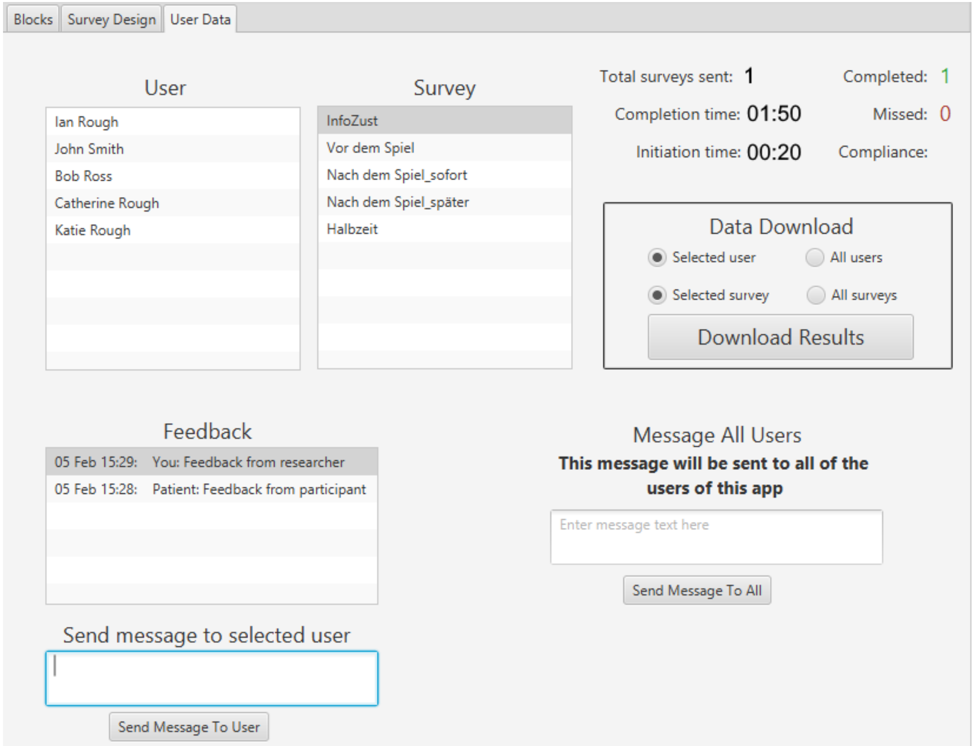


Figure 5.8: The Jeeves User Data Pane, implementing Requirements 8 and 9

greater detail in Section 5.5.

Figure 5.8 shows the User Data pane implemented in Jeeves to let researchers view and interact with participant data. The upper half of the screen displays participant data on survey compliance, as well as initiation and completion times. By selecting a participant and a survey, data can be downloaded in a spreadsheet format for all, or selected participants and surveys. ESM data can be analysed in a number of different ways, as described in Chapter 2, and so to a limited extent, researchers can choose the type of data required to address their research questions.

Requirement 9 - Jeeves should support communication with participants

In general, ESM research studies demand a level of scientific rigour, such that it may be inappropriate to contact or communicate with participants through the duration of a study. However, directly contacting participants to prompt their compliance, and allowing participants themselves to contact the researcher with feedback, have been cited as useful features in psychology research [160]. This functionality has even greater utility in clinical applications, where patients desire a means to ask questions of their healthcare provider, and in turn, these providers may desire a way to easily provide personalised feedback to patients [98].

Further, two-way communication is an important condition for enabling effective meta-design.

Table 5.5: A summary of design requirements for EUD-ESM tools

Category	Design Requirement	Necessity
Sampling Strategies	Support signal, interval, and event-contingent sampling	Must
	Support context-sensitive sampling	**Should**
Participant Tailoring	Support tailoring to individual participants	Must
	Allow researchers to define the level of tailoring	**Should**
Assessment and Intervention	Support all question types used in paper ESM studies	Must
	Support conditional branching based on question responses	Must
	Support tailored interventions	**Should**
Communication	Let researchers monitor participant data in real-time	Must
	Support communication with participants	**Should**

As discussed in Chapter 4, the meta-design framework supports software developers, end-users, and other stakeholders, to actively communicate and collaborate in the refinement of software. For example, pilot study participants could communicate to a researcher that too many surveys are being sent, allowing studies to be progressively refined during their test deployment. This resulted in the design implication of Chapter 3 that **a means of direct feedback between researcher and participant should be incorporated**. The lower half of the User Data pane in Figure 5.8 satisfies this requirement. Jeeves contains a messaging widget that allows two-way communication between participants. Additionally, the widget on the lower-right enables messages to be sent to all participants simultaneously.

5.2.5 Summary

These nine design requirements, summarised in Table 5.5, are not intended to be comprehensive requirements for an EUD-ESM tool. However, having been derived from the literature of ESM in Chapter 2, and the design implications established in Chapters 3 and 4, they represent the synthesis of literature reviewed in this thesis. Indeed, these requirements address the four desirable features derived from the interaction model of Chapter 3, thereby addressing the design implication that **the four features derived from Chapter 3 should be implemented in Jeeves**. Further, Table 5.6 shows to what extent these specific requirements are addressed in the tools referenced in Chapter 4. The following section describes decisions regarding visual language design that enables these requirements to be employed by non-programmers.

Table 5.6: Design requirements specified in Section 5.2 and their fulfillment in EUD-ESM tools

	iPromptU[1]	PsychLog [150]	PACO[2]	AWARE [151]	Ohmage [152]	Sensus [153]	PartS [101]	SurveySignal[3]	LifeData[4]	MovisensXS[5]	EthicaData[6]	Jeeves
R1 - Standard Sampling Strategies	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R2 - Context-Sensitive Sampling	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓
R3 - Tailoring to individual	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✓
R4 - Researcher -defined tailoring	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
R5 - Various survey question types	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
R6 - Conditional question branching	✗	✗	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓
R7 - Variable interventions	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
R2 - Real-time data monitoring	✗	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
R9 - Communication with participants	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓

5.3 Visual Syntax and Semantics

The syntax and semantics of visual programming languages are often implemented with intentions of being *easy to use* or *user-friendly*, but are less frequently based on concrete design guidelines. For example, many languages are implementations of the Blockly library[7], but even this structured, blocks-based paradigm allows for a diversity in syntax and semantics. Blockly-based environments frequently do not justify why the decision was made to use this approach, beyond making their particular domain more accessible to end-users. Hence, although Chapter

[1] http://www.cogtherapy.com/ipromptu.htm. Accessed 17/03/18
[2] http://www.pacoapp.com. Accessed 17/03/18
[3] http://www.surveysignal.com. Accessed 17/03/18
[4] http://www.lifedatacorp.com. Accessed 17/03/18
[5] https://xs.movisens.com. Accessed 17/03/18
[6] http://www.ethicadata.com. Accessed 17/03/18
[7] https://developers.google.com/blockly/ Accessed 17/03/18

4 proposed the design implication that **a blocks-based programming paradigm should be employed in Jeeves to provide flexibility to non-programmer developers**, specific elements of this implementation necessitate more specific design guidelines.

If such guidelines are not considered, even the blocks-based approach can suffer from usability issues. In order to justify the use of Jeeves, two sets of guidelines for evaluation of visual languages are introduced and described, namely the Cognitive Dimensions of Notations [161], and the Physics of Notations [162]. The former is better suited to evaluating interactive artifacts, and is used to guide the *writability* of Jeeves. The latter does not take into account interactive qualities, but instead provides useful, empirically derived principles for enhancing the *readability* of Jeeves. Although both sets of guidelines have structured the initial design process, feedback from usability and case studies in later chapters has influenced design features in ways that conflict with certain guidelines.

5.3.1 Writability - Cognitive Dimensions

The Cognitive Dimensions of Notations were introduced by Green in 1989 as criteria for assessing the usability of both interactive user interfaces and non-interactive notations [163]. Consisting of a small set of concepts that are relevant to any "information artifact", they have previously been applied in detail to the specific domain of visual programming languages [161]. The Cognitive Dimensions[8] serve as discussion tools, not as absolute qualities of a desirable system, such that different systems will require different positions on the axis of each dimension, with some dimensions trading off against others.

The Cognitive Dimensions are used in this section to structure the justification of the various decisions taken in the design of Jeeves, and the trade-offs in other dimensions that these decisions have introduced. The dimensions of interest, and their relationships to other dimensions, are summarised in Table 5.7.

5.3.1.1 Closeness of mapping: What 'programming games' need to be learned?

The closeness-of-mapping dimension defines how closely problem entities in a domain are mapped onto task-specific program entities. In the problem domain of ESM, programmatic representations are inherently event-driven; for example, "**when** something happens (an appropriate time, or a button press), **do** something else (send a survey, send a prompt)" is an example of an event-driven programming statement.

[8]They will hereby be referred to as "Cognitive Dimensions" for the sake of brevity

Table 5.7: Strong dimensions of Jeeves, and their positive and negative influence on other dimensions

Strong Dimension	Positively affects	Negatively affects
Closeness of mapping	Role-expressiveness	Diffuseness Abstraction
Error-proneness	Hidden dependencies Viscosity	N/A
Premature commitment	Viscosity Error-proneness	N/A
Secondary notation	Role-expressiveness Consistency Error-proneness	N/A
Viscosity	Premature commitment	N/A
Visibility	Viscosity Progressive evaluation	Hidden dependencies Abstraction
Consistency	Error-proneness	Hidden dependencies Hard mental operations

Blocks-based programming languages support a one-to-one mapping between problem concepts and task-specific program entities (in this case, graphical blocks) provided that these concepts are carefully chosen. Further, studies of Scratch and App Inventor show that students can quickly create complex apps with an event-driven structure, suggesting that blocks are appropriate for mapping problems that can be expressed as such.

In Jeeves, the event-based problem concepts have been derived from previous ESM studies, including relevant triggers, actions and conditions. Mapping these concepts to blocks, and grouping these blocks by the type of function they perform, appears to simultaneously increase the **role-expressiveness** of the notation, defined as how well a visual representation of elements communicates their purpose.

Trade-offs?

The problem concepts defined in Jeeves do not enable additional **abstraction** beyond that introduced in the design. Blocks can be combined to provide the functionality of an app, but combinations cannot be encapsulated into larger abstract blocks. Allowing flexibility in abstraction is not necessarily a positive feature, as time must be taken by end-user developers to learn and create these additional abstractions.

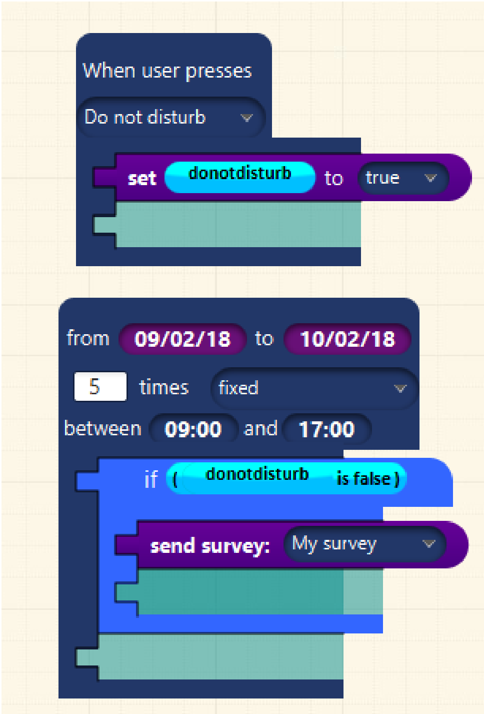


Figure 5.9: Diffuse representation of a 'Do not Disturb' button

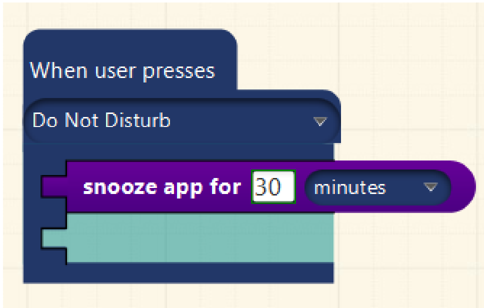


Figure 5.10: More terse representation, but with a program entity not mapped to a derived problem concept

A further trade-off to mapping only concepts considered relevant to a problem is that of **diffuseness**, or how many entities are required to express a meaning. Consider Figure 5.9, which shows the design of a 'Do Not Disturb' button for restricting survey prompts, and Figure 5.10, a considerably more **terse** example of the same function. The 'snooze' action was not initially felt to map closely to the problem domain of experience sampling. However, as a result of a usability study described in Chapter 6, the snooze action is now an adopted block. Further studies may trade off closeness-of-mapping for less diffuse representations in a similar way.

5.3.1.2 Error-proneness: Does the design of the notation induce 'careless mistakes'?

Visual languages are well-suited to reducing error-proneness in novice users. Particularly in Jeeves and similar blocks-based languages, the puzzle-shaped program entities afford syntactically correct combinations, which appear to visually fit together. When one attempts to insert a block where it does not make syntactic sense, the block does not appear to 'snap' into place, and instead sits detached on the canvas. The reduction of syntactic errors by definition improves app quality and reliability - of particular importance in EUD where the developed artifact is to be used by a separate group of users (public-outward EUD [142]). Minimising the opportunity for errors has a generally positive effect on developer experience [138], but particularly in relation to the Cognitive Dimensions of **hidden dependencies** and **viscosity**.

The puzzle-piece block shapes, which suggest how they should fit together, makes hidden dependencies between programmatic constructs visually explicit. Further, a notation that does not allow for erroneous changes increases the ease with which correct changes can be made, thereby reducing the viscosity of the notation.

Trade-offs?

There do not appear to be any direct trade-offs to reducing error-proneness in Jeeves, or indeed any language. Although there are *indirect* trade-offs such as the visual verbosity of the blocks paradigm, it is never desirable for a language to be far along the error-proneness dimension.

5.3.1.3 Premature commitment: Do programmers have to make decisions before they have the information they need?

Green and Petre have identified different types of commitment that languages may enforce [161]. For example, some languages have a *layout commitment*, where developers must commit to placing visual components in certain places on-screen. This is avoided in Jeeves by the ability to freely move blocks around at any point. Further, separate triggers do not need to be joined together, removing the potential for ‘spaghetti diagrams’ that can cause problems in flow-based visual languages. Jeeves also avoids a *connection commitment* - if an action is found to be in the wrong trigger, for example, then it can be moved to the correct trigger in one drag-and-drop motion, ensuring **viscosity** is reduced. One potential issue identified in Jeeves is the *commitment to construct* - if multiple actions are found to be in the wrong trigger, then there is no way to transfer them as a group; each must be moved individually. The magnitude of this issue is reduced through ensuring that Jeeves enforces no *commitment to order of creation*. Developers can identify appropriate actions before knowing which trigger to use, or define an attribute before knowing which expression it will belong in.

Commitment to order of creation is prevalent in EUD-ESM tools based on Web forms. Although forms are a familiar input mechanism for the majority of computer users, and a common paradigm for EUD-ESM, the guided, linear approach could be problematic. In usability studies conducted in Chapter 6, participants would often define actions before deciding how to trigger them. Others would create triggers, then fill them with actions. A researcher’s mental model may be inhibited by forcing premature commitment to either [161]. Instead, blocks can support researchers in translating their study protocols in a *bricolage* fashion [164], thereby reducing **error-proneness**.

Trade-offs?

As with error-proneness, while the conditions affording low premature commitment may have their own disadvantages, there can be no *direct* trade-off to ensuring developers have the knowledge necessary to avoid programming incorrect functionality.

5.3.1.4 Secondary notation: Can programmers use layout, colour, other cues to convey extra meaning, above and beyond the ‘official’ semantics of the language?

In virtually all domains and languages, visual or otherwise, the use of secondary notation is beneficial. Although Jeeves in its current state does not allow comments or annotations, other sources of secondary notation have been used to facilitate construction. For example, the shape and colour of blocks are suggestive of where they should be integrated into the specification, reducing the potential **error-proneness** caused by dragging incompatible blocks together. Further, the canvas of Jeeves allows users to arrange triggers as they see fit, potentially grouping them by functionality or ordering time-based triggers chronologically, to employ their own secondary notation.

In Bertin’s *Semiology of Graphics*, eight visual variables are defined that can be used to distinguish information graphically [165]. These are: horizontal position, vertical position, shape, colour, size, brightness, texture, and orientation. Jeeves makes direct use of the two positional variables, as well as shape, colour and size. While not making full use of potential information, this secondary notation is adequate to ensure the **role-expressiveness** of each block. When the role of each block type is established through this secondary notation, it simultaneously enhances the **consistency** of Jeeves as developers learn to apply the same block patterns.

Trade-offs?

Again, there are no direct trade-offs to employing secondary notation. As expressed in Green and Blackwell’s tutorial on the Cognitive Dimensions: *“It is hard for me to imagine any situation that would not be improved by making a notation easier to read”* [166, p. 29].

5.3.1.5 Viscosity: How much effort is required to perform a single change?

A low viscosity is a major factor in the design of Jeeves, in order to ensure that changes can be made easily, particularly by novice users. The blocks paradigm supports this; adding, removing, or swapping an action in Jeeves is a simple case of dragging and dropping. Actions and expressions can be swapped about in their triggers, moved to different triggers, or removed entirely without the need to reconnect wires or rearrange boxes, as in flow-based visual notations. A low viscosity also has a mutually positive effect on **premature commitment**. By ensuring a study is simple to modify, developers are not committed to a specific design, as any changes can be easily reversed.

Trade-offs?

Green and Blackwell explain that high viscosity can sometimes be beneficial when dealing with safety-critical systems that could have disastrous consequences through small modifications [166]. While the magnitude of consequences are dependent on its context of use, appropriate role access is a more suitable safety mechanism than introducing unnecessary resistance to change.

5.3.1.6 Visibility: Is every part of the code simultaneously visible, or is it at least possible to juxtapose any two parts side-by-side at will?

The visibility of study specifications in Jeeves is high in comparison to many alternative EUD-ESM tools. Those that are composed of Web forms require navigation across separate pages, making it difficult to see an app's function in its entirety. In contrast, the canvas of Jeeves, which supports panning and zooming operations, ensures that even complex applications can be viewed in one screen. Surveys and user data can also be viewed concurrently with the blocks if required, through adjustable panels.

Visibility in Jeeves supports **progressive evaluation** - with the entire study specification visible, it should be simpler to evaluate how an app's functionality changes with respect to small changes in the blocks. This likewise has a positive affect on **viscosity** - if an app malfunctions due to an incorrect specification, the error can be located on-screen, without the need to navigate through a hierarchy of menus.

Trade-offs?

High visibility does, however, introduce trade-offs. In ensuring that all blocks are visible on the canvas, Jeeves does not allow developers to group triggers and actions together into their own **abstractions**, a feature that was suggested by usability study participants in Chapter 6. Further, the visibility of Jeeves ironically introduces **hidden dependencies**. For example, explicitly showing dependencies between survey questions that set attribute values, and these attribute blocks on the canvas, would introduce extra visual clutter that would sacrifice overall visibility of the study specification.

5.3.1.7 Consistency: When some of the language has been learnt, how much of the rest can be inferred?

The visual language of Jeeves is designed for consistency; secondary notation ensures that triggers, actions and expressions are clearly identified, and all blocks within a given category have the same behaviour. This consistency of Jeeves minimises **error-proneness**. Once the general concepts of dragging actions into conditions/triggers, building surveys, and assigning

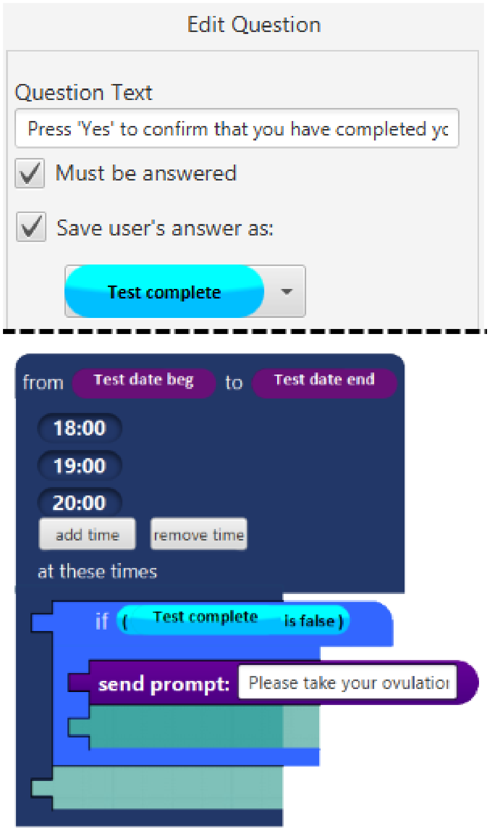


Figure 5.11: Participants had to work around the lack of a 'survey expression'

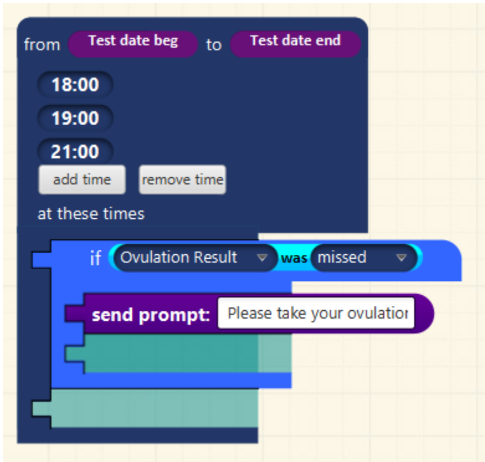


Figure 5.12: The 'survey expression' makes implementation simpler, but sacrifices consistency

attributes through surveys are understood, there are no new rules or combinations of actions that could induce further errors.

Trade-offs?

While the design of Jeeves was not expected to involve **hard mental operations**, user studies identified occasions where this was the case. For example, Figure 5.11 shows how a case study user implemented a function to prompt a user repeatedly if they had not completed a survey. The user had to find a somewhat complicated workaround by setting the value of a $testcomplete$ attribute in an additional survey question, which was used in the blocks specification to determine whether the survey had been completed. Figure 5.12 shows how this can be resolved through a "Survey Expression" that returns whether or not a participant has completed a particular survey that day. However, a general syntactic rule in Jeeves is that all expressions contain at least one attribute. Thus, this expression introduces an inconsistency; there is a clear trade-off here.

The goal of consistency has introduced further **hidden dependencies**. Attribute values are consistently set through survey questions, whether they are participant preferences or attribute values that trigger interventions. If the developer does not ensure that these preferences are established through surveys at the beginning of the study, this could cause errors. Consider the triggers in Figures 5.11 and 5.12. If the date attributes *Testdatebeg* and *Testdateend* are not initially set, then the trigger will never fire. An inconsistent approach could have a separate means of establishing participant preferences that would minimise the chance of the developer forgetting to do so.

5.3.2 Readability - Physics of Notations

While the Cognitive Dimensions have been applied effectively in the evaluation of visual programming languages, applying this language in practice is an *encoding* issue (related to the developer's construction of the visual program). With respect to *decoding* issues, it is important to ensure that programmatic representations are readable by those who did not create them. This is a particularly important issue in Jeeves for a number of reasons.

- Researchers and clinicians are likely to work in teams. It is important that a representation created by one team member is comprehensible by another.
- As well as intra-team readability, *inter*-team readability is important. Scientific research is facing a "replicability crisis" [15], which formal representations of ESM study methodologies could help address.
- Researchers, and clinicians in particular, may collaborate with participants to design an appropriate app. A visual representation could support shared understanding.

Previous work has shown that natural language is unsuitable for making the distinction between triggers (discrete events) and conditions (ongoing states) [167]. Even in a blocks-based language, triggers and actions can be confused if they are not distinct [168]. This distinction may not be difficult during the *encoding* stage, but is likely to be misunderstood during *decoding*. It is therefore important that the intentions of one developer are easily interpreted by another.

Readability aspects of the notation employed in Jeeves are described in terms of Moody's Physics of Notations, which synthesises design principles for enhancing the *"cognitive effectiveness"* of a visual notation. Cognitive effectiveness is defined as the speed, ease, and accuracy with which a representation can be processed by the human mind [169]. Each of the nine principles is briefly described, and how the design of Jeeves conforms, or does not conform, to each.

5.3.2.1 Principle of Semiotic Clarity

This principle states that *"There should be a 1:1 correspondence between semantic constructs and graphical symbols"*, which is a prescriptive definition of the *closeness-of-mapping* Cognitive

Dimension. Instead, Jeeves contains a “symbol deficit” in that graphical symbols are not used for surveys or sensors, which are instead represented textually. Thus, this principle is only partially applied with respect to triggers, actions and conditions.

5.3.2.2 Principle of Perceptual Discriminability

Moody’s second principle is that *“Different symbols should be clearly distinguishable from each other”*, similar to the *role-expressiveness* Cognitive Dimension. A visual notation where different types of symbol are distinct in appearance, as in Jeeves, is strong in role-expressiveness and abides by this principle.

5.3.2.3 Principle of Semantic Transparency

This principle states that notations should *“Use visual representations whose appearance suggests their meaning”*. Blocks in Jeeves are used to represent abstract, event-driven concepts such as triggers, actions and conditions, as opposed to tangible objects. They are *semantically opaque*, in that their visual appearance is arbitrarily related to their meaning. However, semantic transparency can also apply to representations of relationships between symbols. The puzzle-piece block shapes are used to transparently communicate how they should be combined. The user attributes shown in Table 5.3 (p. 94) also make use of secondary notation to distinguish their relationships with other blocks. For example, time and date attributes are identical in shape, colour and size to the date and time pickers on triggers, implying their relationship before it is explicitly known. While the *Principle of Perceptual Discriminability* declares that symbols should be visually distinct, this *Principle of Semantic Transparency* declares that their visual distinction should also communicate their relationship.

5.3.2.4 Principle of Visual Expressiveness

This principle, *“Use the full range and capacities of visual variables”*, acts as a prescriptive definition of the Cognitive Dimension of *secondary notation*. Moody references Bertin’s eight visual variables [165], to explain what kinds of secondary notation are possible, and how they should be used. The two extremes of a notation’s visual expressiveness are *non-visual* (zero variables utilised) and *visually saturated* (all eight variables utilised). In using five visual variables, as discussed in the previous section on Cognitive Dimensions, Jeeves provides sufficient semantic meaning to study logic. Brightness, texture and orientation could be incorporated in future if the need arises for further distinctions.

5.3.2.5 Principle of Dual Coding

This principle, *"Use Text to Complement Graphics"*, emphasises the importance of textual representation where visual syntax ends up being more cumbersome and difficult to manage. There is no direct mapping of this principle to a Cognitive Dimension, although that of *secondary notation* describes the possible use of annotations and comments.

Moody claims that *"text should never be used as the sole basis for distinguishing between symbols"* [162, p. 771]. However, in Jeeves, text and customisation widgets are the only means by which blocks in the same category are distinguished from one another. In many visual languages, the "graphics-text boundary" is shifted in favour of graphics accompanied by minimal text (for example, *LabVIEW*[9]). In Jeeves, this boundary balances the two means of encoding equally. Figure 5.13 illustrates Moody's representation of three different encoding strategies distinguished by whether they are graphical or non-graphical, and on-diagram or off-diagram. According to dual coding theory [170], using text and graphics together to convey information is more effective than using either on their own. This is a strong argument for representing an ESM study with blocks, where the visual and text representations complement one another, unlike the vast majority of other EUD-ESM tools.

5.3.2.6 Principle of Graphic Economy

Research suggests that the human ability to discriminate between perceptually distinct alternatives, the *span of absolute judgment*, is around seven categories [171], providing a strong argument for this principle, that *"The Number of Different Graphical Symbols Should Be Cognitively Manageable"*. The visual notation in Jeeves represents triggers, actions, conditions, expressions, and attributes. Prior to this design, additional graphical symbols were used to represent sensors and surveys. However, these concepts are not related to the logical structure of a study, and are thus represented otherwise: sensors are encoded as text (Figure 5.13, *Textual Encoding*) and surveys are represented off-diagram in the Survey Design pane (Figure 5.13, *Supporting Definitions*).

This principle does not account for the quantity of symbols on a diagram, hence it is not directly related to visibility. However, introducing symbol deficit by encoding some information *off-diagram* does indeed affect visibility when diagrammatic and non-diagrammatic information must be juxtaposed.

[9] http://www.ni.com/en-gb/shop/labview.html Accessed 17/03/18

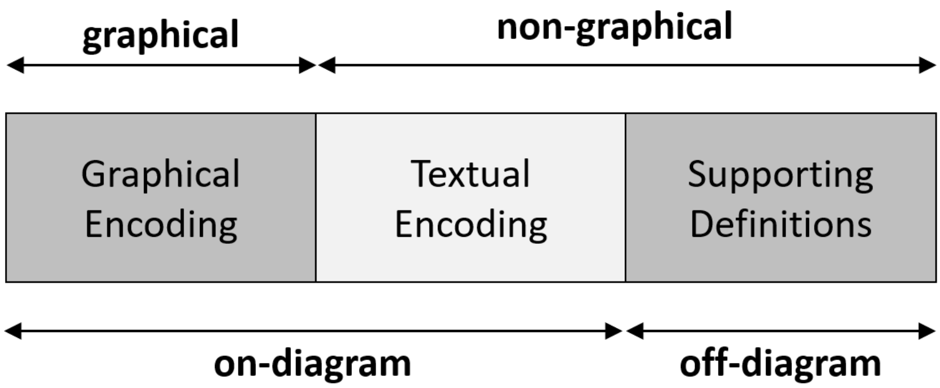


Figure 5.13: Different means of encoding information (reproduced from [162], ©2009 IEEE)

5.3.3 Design principles not applied

Three design principles from the Physics of Notations are not applied in the current design of Jeeves, but instead could guide future research.

5.3.3.1 Principle of Complexity Management

This principle states that notations should *"Include explicit mechanisms for dealing with complexity"*. Discriminating between diagram elements becomes more difficult with increasing diagram size, with a consequence that visual representations of large problems may cause cognitive overload. Alternative EUD-ESM tools surveyed in Chapter 4 manage complexity through a hierarchical structure of Web forms for different triggers and actions. However, the visibility of such an approach is comparatively poor, requiring a series of navigation steps to obtain an overview of a study. Moody suggests the use of encapsulation as a form of complexity management, such that this principle is directly related to the Cognitive Dimension of *abstraction*. In contrast, Jeeves has no way of managing complexity through abstraction, with a focus on maximising visibility.

5.3.3.2 Principle of Cognitive Fit

This principle, *"Use different visual dialects for different tasks and audiences"*, has also not been applied. There is no difference in how novices and experts read or create a study specification in Jeeves. While the usability evaluations in Chapter 6 suggest that the notation is understandable by users with any level of programming experience, it remains to be investigated whether a different notation would be appropriate for a *reader* as opposed to a *writer*, and whether programmers or non-programmers would differ in their preferences.

Table 5.8: Principles in relation to Cognitive Dimensions. ✓= Positively related, ✗= Negatively related

Principle	Related Cognitive Dimension	Principle applied in Jeeves?
Semiotic Clarity	Closeness-of-mapping ✓ Diffuseness ✗	✓✗
Perceptual Discriminability	Role-expressiveness ✓ Secondary notation ✓	✓
Semantic Transparency	Role-expressiveness ✓ Secondary notation ✓ Hidden dependencies ✓	✓
Complexity Management	Abstraction ✓ Visibility ✗	✗
Cognitive Integration	Abstraction ✓ Visibility ✗	✗
Visual Expressiveness	Role-expressiveness ✓ Secondary Notation ✓	✓
Dual Coding	Role-expressiveness ✓ Secondary Notation ✓	✓
Graphic Economy	Diffuseness ✓ Secondary Notation ✓ Role-expressiveness ✓ Closeness-of-mapping ✗ Visibility ✗	✓✗
Cognitive Fit	Abstraction ✓ Visibility ✗	✗

5.3.3.3 Principe of Cognitive Integration

At present there is no way in Jeeves to *"Include explicit mechanisms to support integration of information from different diagrams"*, as declared in this principle. Each study specification exists as a standalone project that can be loaded, modified and saved, but not combined. Future work could investigate the utility of modularising entire specifications that could then be integrated into more complex combinations.

5.3.4 Summary

Table 5.8 summarises the relationships identified between the Physics of Notations and the Cognitive Dimensions. While it is not clear whether these relationships apply to all visual notations, future work could explore how well they apply to blocks-based programming notations in general, such as those implemented with the *Blockly* library. There are clear parallels between the two notations; in particular, all the principles succeed in prescribing ideal magnitudes of particular Cognitive Dimensions, or combinations of them. Cognitive Dimensions not addressed in the Physics of Notations are: consistency, error-proneness, hard mental operations, premature commitment, progressive evaluation, and viscosity. These dimensions appear to be exclusively relevant to the *encoding* process.

5.3.4.1 Dimensional relationships

In constructing Table 5.8, it was noted that pairs of Cognitive Dimensions had consistent relationships with one another in relation to the different Physics of Notations principles.

Role-expressiveness vs. Secondary notation: It was difficult to distinguish between *role-expressiveness* and *secondary notation* in relation to five of Moody's principles. In the encoding process, the roles of blocks are primarily distinguished by the categorical menu in which they are found. For example, a developer looking for a trigger block will find it in the *Triggers* menu. However, this form of textual role-expressiveness is absent when reading a specification, making it important to maximise the use of secondary notation to distinguish the role of each block.

Closeness-of-mapping vs. Diffuseness: Additionally, there was a difficulty distinguishing *closeness-of-mapping* and *diffuseness*, which are negatively related to one another in the two conflicting principles of Semiotic Clarity and Graphic Economy. The *Principle of Semiotic Clarity* states there should be a 1:1 mapping between symbols and concepts, whereas the *Principle of Graphic Economy* states that symbol deficit may be required to reduce cognitive load. Consequently, if a large number of problem concepts must be mapped, then these principles must clash. Following usability and case studies with Jeeves, additional symbols have been added in response to user requests. If Jeeves continues to grow in the number of concepts it must represent, then abstraction will have to be implemented to reduce diffuseness. Graphic Economy and Semiotic Clarity are both partially implemented in Jeeves as previously discussed, and indicated by ✓✗ in Table 5.8.

Abstraction vs. Visibility: *Abstraction* also has a strong relationship with *Visibility*. These Cognitive Dimensions are negatively related within three of the principles. For example, abstraction could be implemented at the expense of visibility, or eliminated to maximise visibility.

Regardless of which is chosen, the three relevant principles (complexity management, cognitive integration, and cognitive fit) will not clash with one another. However, these principles, as shown in Table 5.8, have not been applied in Jeeves, as they all require the ability for developers to understand the process of abstraction creation.

5.4 Android Application

Although the focus of this thesis is on factors influencing usability and utility from the researcher's perspective, it is nevertheless important to address considerations for design of the corresponding Android app. The overarching goal of the JeevesAndroid app is to minimise participant burden, and design decisions have therefore been made in pursuit of this goal.

5.4.1 Overview

Participants in Jeeves studies use a native Android app referred to as *JeevesAndroid*, which downloads a study specification from an online database, the details of which are described in the next section. JeevesAndroid dynamically interprets this specification, setting up the necessary triggers, actions and surveys automatically, such that studies can be updated in real-time. Survey responses and participant feedback are similarly uploaded to the database in real-time whenever an Internet connection is available. While web apps are compatible with different smartphone operating systems, a native app provides access to all bespoke features of the device, increasing the flexibility of data capture and triggering possibilities. The utility of Jeeves could be increased by having an equivalent iOS app, enabling a greater number of participants to be recruited. However, for the purposes of this thesis, the focus is on the Android client, while future work could include an iOS implementation.

Implementing a native Android app has also enabled use of the *EmotionSense* framework - a set of open source libraries for sensor data capture, classification, and trigger scheduling [172]. JeevesAndroid integrates the *Sensor Manager* and *Trigger Manager* libraries, which have simplified the implementation of context-aware functionality. The *Sensor Manager* provides a consistent API to sample data from a range of available sensors, at specified intervals and granularity. The *Trigger Manager* uses the specified constraints of the particular trigger events, as well as stored characteristics and preferences of participants themselves, to determine whether to execute particular actions. By combining the two components, sensor-based triggers can also be implemented.

However, maximising compliance in both installing and completing surveys with JeevesAndroid is dependent on minimising participants' burden in doing so. Compliance and attrition are

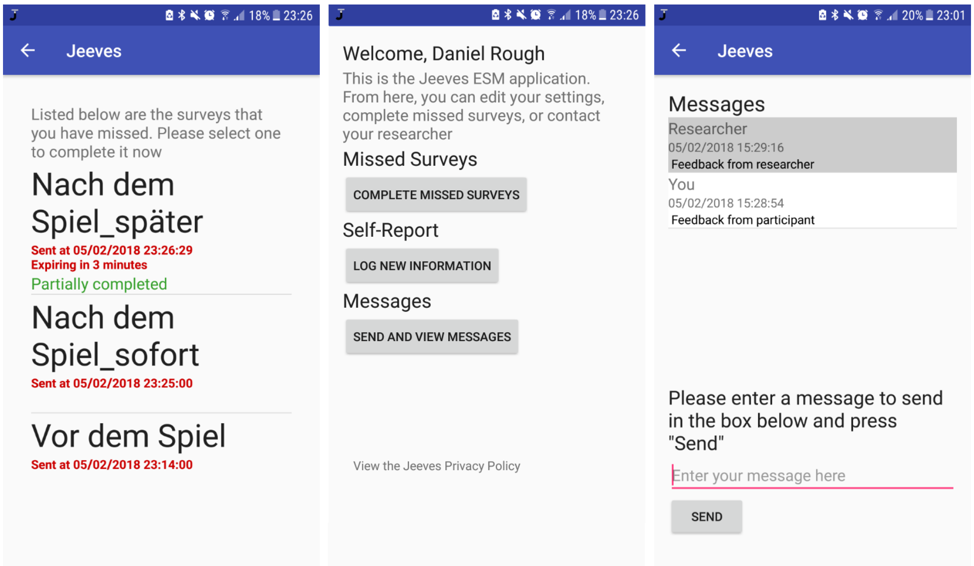


Figure 5.14: Left: Missed surveys screen. Centre: Main app screen. Right: Feedback screen

salient issues in the majority of experience sampling studies, and have been the primary factors influencing the following design decisions, related to efficiency, security, and battery life.

5.4.2 Designing for efficiency

As explained in Nielsen's usability heuristics: *"Every extraneous unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility"* [173, p. 339]. With respect to this, JeevesAndroid is designed with a minimalist interface to increase efficiency. The main app screen, after a participant registers with a particular study, presents a simple choice of three buttons (Figure 5.14, center).

By tapping the "Complete Missed Surveys" button, participants can see a list of surveys that have not been completed, but have not yet expired, as shown in the leftmost screenshot of Figure 5.14. By tapping on a survey in the list, a participant can either begin the survey if it has not been initiated, or continue from their last answered question. For example, the top survey in the screenshot shows *Partially completed* to indicate that the participant has already entered answers for some questions. In summary, the goals were to provide simple navigation to postponed surveys, and to ensure that previously entered data did not have to be re-entered, thereby increasing participants' efficiency.

The rightmost screenshot of Figure 5.14 shows the participant feedback screen. This section of JeevesAndroid allows participants to send and receive messages to and from researchers administering a study. Although delivery of personalised feedback to participants is at the

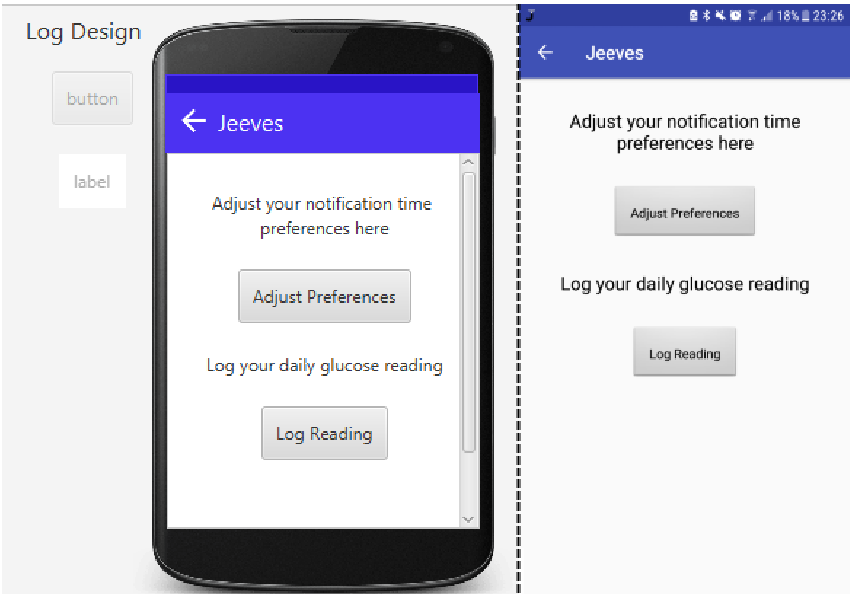


Figure 5.15: Left: Log Design pane of Jeeves. Right: Corresponding self-report screen on JeevesAndroid

discretion of researchers in their study goals, the one-way feedback **from** participants is useful from a meta-design perspective. Participants' feedback during a study could allow researchers to adjust a specification to handle unexpected app behaviour, or request new features from meta-designers if an adjustment cannot be made.

Finally, the "Log New Information" button brings a participant to the screen previously designed by the researcher, as shown in Figure 5.15, right. Researchers' interface widgets are currently presented on this separate screen, unlike the first implementation of JeevesAndroid where these widgets were appended to those on the welcome screen. Regarding Nielsen's usability heuristic of minimalist design [174], a potentially high number of buttons and labels on the welcome screen could be initially overwhelming to a participant, potentially causing compliance issues.

5.4.3 Designing for battery life

Battery life is a primary concern for an app that will be continuously running in the background of a participant's smartphone, potentially for many days or weeks. During the initial implementation of JeevesAndroid, it became apparent that manually coded functions for recognising activity, and detecting geofences, was inefficient and resource-intensive. However, continuous development of Google's APIs have mitigated this, as described in this section.

5.4.3.1 Geofencing

A popular use of context-awareness in smartphone apps is to initiate actions when a user enters, leaves, or dwells in a particular location. As discussed in Chapter 2, such functionality can be used to gather further, contextually relevant information through a self-report survey, or deliver motivational messages at a triggering location. However, a smartphone's GPS sensor is not designed for continuous sensing, such that even simplistic geofencing capabilities place a high demand on smartphones' resources. To avoid this issue, JeevesAndroid uses Google's Places API[10], which provides a means of setting up geofences that minimise the use of the phone's resources, at the expense of minor delays. Google's classifiers employ a combination of sensors, including WiFi and accelerometer, to minimise resource consumption.

5.4.3.2 Activity recognition

Appropriate use of accurately inferred activity can significantly reduce the burden of an ESM app on a participant. For example, inference that a participant is driving could delay prompts to ensure that no accidents occur. A sudden change in acceleration can imply a transition between activities – an opportune moment to interrupt [84]. Additionally, recognition of particular activities such as running or cycling could prompt for self-report regarding a participant's recent exercise. As with geofencing, however, continuous sensing of accelerometer data can be burdensome on the processing power of a smartphone. Google's Activity Recognition API[11] was utilised in JeevesAndroid, to minimise potential battery drainage and resource consumption. The API employs power-saving strategies such as reducing sampling frequency when a device has been still for an extended period.

5.4.3.3 Data upload

When a network connection is unavailable, continuous attempts to upload data are resource-intensive, which is aggravated by the Jeeves design requirement of real-time synchronicity between researcher and participant. In the initial JeevesAndroid implementation, data upload to a custom-built server was configured manually, causing significant drain on smartphones' battery life. In the current implementation, which uses the *Firebase* online database and API, data is stored locally until a network connection is established, significantly reducing resource consumption.

[10] https://developers.google.com/places/ Accessed 17/03/18
[11] https://developers.google.com/location-context/activity-recognition/ Accessed 17/03/18

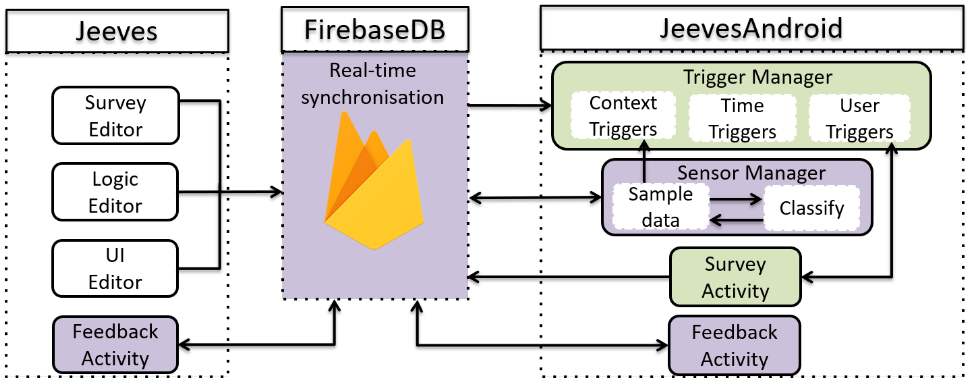


Figure 5.16: Simplified architecture of the Jeeves platform

5.4.4 Designing for security

A design requirement inherent to ESM apps is the security of sensitive participant information in both transfer and storage. Such information includes the participant's personal details, as well as their self-report responses. As one layer of security, the desktop and mobile client operate a public/private key protocol, so that participants' information and their survey responses can only be read by the researcher who designed and launched the study. While other researchers may still be allowed access to the study specification and survey questions, the responses of the participants assigned to that study will be encrypted from an outside perspective.

5.5 Platform Overview

Figure 5.16 illustrates a simplified architecture of the Jeeves platform. In this section, the functionality of Firebase and the semi-structured representation of studies will be briefly introduced. The use of Firebase is justified with regards to the design requirements addressed at the beginning of this chapter.

5.5.1 Firebase

Firebase[12] is a platform for developing mobile apps, which includes a real-time database that allows data to be synchronised across multiple devices on different platforms. Study specifications and participant data are stored in the Firebase online database, and transferred between researcher and participant clients using secure HTTP. Firebase has libraries and API documentation that allow it to be integrated easily into the Android and Java implementations. Features of Firebase that support it as a solution for Jeeves are as follows:

[12] https://firebase.google.com/ Accessed 17/03/18

- **Real-time synchronisation:** Firebase allows survey data uploaded by participants with JeevesAndroid to be delivered to Jeeves instantly. Additionally, any updates to a study made with Jeeves get instantly delivered to JeevesAndroid participants, enabling real-time monitoring and modification.

- **Offline storage:** Participants may not always be online when completing surveys, or when studies have been updated. Firebase allows data to be stored in a participant's smartphone cache to be uploaded, and study updates to be synchronised, as soon as a reliable network connection is available.

- **Semi-structured data:** The Firebase database is a NoSQL database, which stores data in a semi-structured, JSON-like representation. This allows dynamic additions and modifications to be made to Jeeves without major reformatting of the database structure. Dynamic updates support a meta-design approach, where changes are continuously made in response to end-user feedback. Examples of the semi-structured data of a Jeeves study specification, and of a particular participant, are shown in Figure 5.17.

- **Security:** Firebase has a set of security protocols for authorisation and authentication of users. In addition, a declarative security model is provided, allowing developers to customise access based on user identity. As an example use, this could allow particular study specifications to be kept private to a researcher, while allowing others to be public for collaboration.

- **Storage:** Firebase has storage support for images, audio, video and other media. This functionality allows multimedia to be presented to participants, or for participants to capture and upload their own media, such as photographs or audio recordings. Firebase Storage is currently used to present images to participants during surveys, including participant information sheets and consent forms.

In summary, Firebase is ideally suited to a remote ESM study application, and has saved significant work, as opposed to manually implementing a complete, custom server solution. Security, real-time synchronisation, media storage, and offline capabilities are all key benefits that should exist in any EUD-ESM tool's server-side platform. However, the extensibility of Firebase makes it particularly useful with regards to meta-design.

5.5.2 Update procedure

The Jeeves visual specification is synchronised with JeevesAndroid to support real-time study modifications. As with all data in Jeeves, each trigger is stored in a hierarchical JSON-like structure, an example of which is shown in Figure 5.18. Each trigger has its own unique,

Figure 5.17: Semi-structured project data (left) and participant data (right)

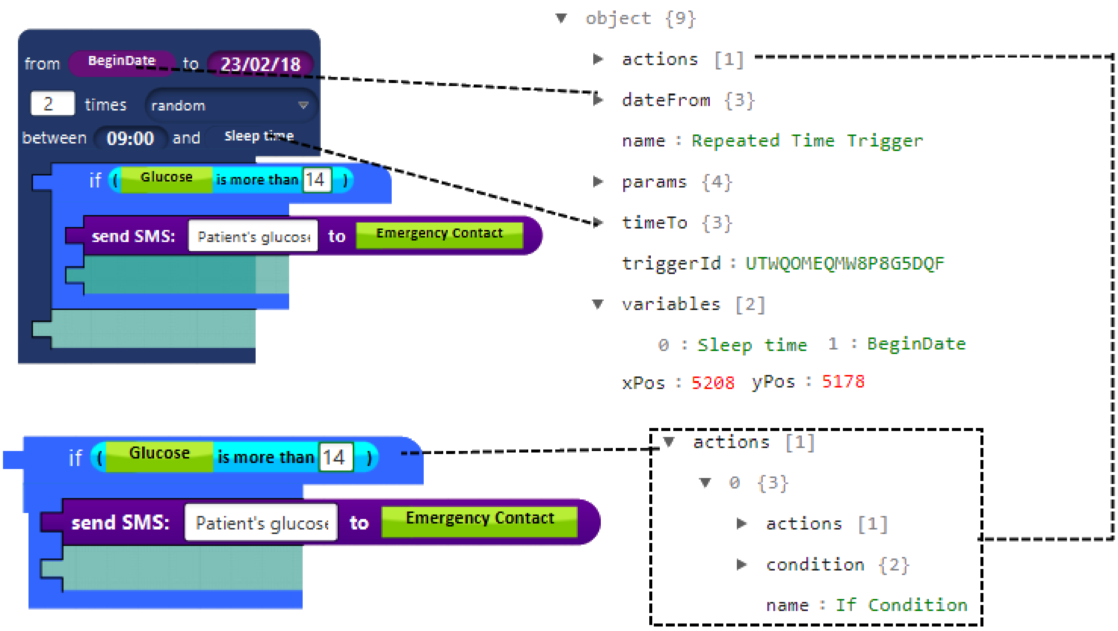


Figure 5.18: Example blocks and their JSON-like structure

randomly generated ID, as well as a list of the attributes it contains (in this case, *BeginDate* and *Sleep time*). This attribute list is required for when a participant updates an attribute value, as will be explained in this section. There are two ways that the logic of a study specification can be updated, by either the researcher or the participant.

5.5.2.1 Researcher updates the trigger hierarchy from Jeeves

1. In the Java implementation of the Jeeves visual programming environment, each block has *listeners* on its various receivers and input widgets. When a single change is made to any receiver or widget in the hierarchy, this change propagates to the trigger's listener, which then generates a new unique ID. For example, in Figure 5.18, changing the conditional value of Glucose from 14 to 15 will generate a new ID for the entire trigger.

2. The Firebase Android API allows listeners to be registered on different parts of the database. In JeevesAndroid, a listener is registered to data of the study to which a participant belongs. For each trigger in the study, the *Trigger Manager* in JeevesAndroid registers a listener on each trigger, and adds it to a map structure, with the Trigger ID as the key.

3. Whenever the study listener detects a change, JeevesAndroid compares the study's new list of trigger IDs to its current list, removing old listeners and adding new listeners as necessary. Thus, for simplicity, a completely new trigger listener is generated for the updated trigger specification.

5.5.2.2 Participant updates an attribute from JeevesAndroid

1. When a participant updates an attribute value, this can change how triggers will behave. In some cases, the value of the attribute is only used when the trigger fires (for example, in the trigger in Figure 5.18, the value of *Glucose* will be checked twice a day at random, as per the trigger specification). However, if the participant updates *BeginDate* or *Sleep time*, the trigger's function is dependent on these new values.

2. In JeevesAndroid, all attributes are stored in Android's *Shared Preferences*, a set of persistent key-value pairs, where the key is the attribute name, and the value is the attribute's current value. An additional listener is added to the set of Shared Preferences to check for any value updates made by the participant.

3. When an update is detected, the listener iterates through the app's list of triggers. If any trigger is found that has functionality dependent on an updated attribute value, then a new

listener is generated for this trigger. For example, if a participant updates the *Sleep time* attribute, then this will create a new listener on the updated trigger.

5.6 Discussion

The implementation of Jeeves represents the synthesis of literature related to experience sampling, and guidelines for visual programming language design. Jeeves therefore serves as a tool that can be used to answer the research sub-questions introduced in Chapter 1. As a visual notation representing an ESM study, Jeeves can be used to communicate ideas to researchers and clinicians, and to understand the usefulness of such a tool, without need for interaction. Further, as a visual programming language that can be *written* as well as *read*, factors related to usability can also be evaluated. Finally, as a full platform consisting of the Jeeves visual programming environment, JeevesAndroid app, and surrounding architecture, it can be evaluated by researchers in practical applications outside the constraints of a usability study.

5.7 Chapter Summary

This chapter discussed the design decisions that were taken into account during the implementation of the Jeeves platform, with particular focus on the visual programming environment. From analysis of existing EUD-ESM tools, as well as recent trends in psychology and medicine, a number of requirements for future EUD-ESM tools were derived. These requirements were described in relation to the blocks-based programming paradigm, the design of which was further described in relation to the Cognitive Dimensions of Notations, and the Physics of Notations. Finally, an overview of the Android app and overall platform architecture was provided.

While design principles derived from previous work, or those already empirically validated, can guide the creation of software that is likely to be usable and learnable by its stakeholders, there is no guarantee that this will be the case. Chapter 6 describes the formal usability evaluations that have therefore been conducted with the Jeeves implementation described.

CHAPTER SIX

USABILITY ANALYSIS

Ideas in this chapter have been published in the following papers:

D. Rough and A. Quigley, "Jeeves - a visual programming environment for mobile experience sampling," in *Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on*, pp. 121–129, IEEE, 2015.

D. Rough and A. Quigley, "Jeeves - an experience sampling study creation tool," *BCS Health Informatics Scotland (HIS)*, 2017.

Perceived ease-of-use is a key factor in the acceptance of technology, exemplified by models introduced in Chapter 1. Furthermore, sets of guidelines for future experience sampling applications advocate allowing researchers to easily create and modify these studies themselves, independent of professional programmers. However, none provide specific details on how this could be achieved. Conversely, a wealth of usability research exists on different visual programming paradigms, including diagrammatic and blocks-based programming, as well as end-user development in general. However, the usability of a programming paradigm, and its encompassing development environment, is heavily dependent on the domain in which it is applied. This chapter thus aims to address the research sub-question related to the perceived ease-of-use of the Jeeves environment, specifically: "*How can the development effort be reduced to allow rapid creation of experience sampling apps without requiring programming experience?*"

6.1 Research Methods

Publications describing EUD-ESM tools generally lack evaluation of both the client-side mobile application, and in particular, the EUD-ESM tool itself. The categories of research purposes proposed by Kjeldskov and Graham [17] and described in Section 1.3, are repeated as follows.

- **Understanding:** Research aimed at understanding the particulars of a phenomenon studied
- **Engineering:** Research aimed at the original development of a tool or technology
- **Re-engineering:** Research aimed at the engineering of modifications or extensions to an existing tool or technology
- **Evaluating:** Research aimed at the assessment, validation and assurance of tools, technology, models and frameworks
- **Describing:** Research aimed at describing the ideal properties of a system or situation

it is notable that the majority of EUD-ESM research is exclusive to the *Engineering* category. In such research, the development of an EUD-ESM tool is described, but with no formal evaluation. It was expected that more systems would additionally be included in the *Evaluating* category, given that 46% of the surveyed EUD publications in [18] which describe a tool also follow up with a lab evaluation. In fact, only two publications report a formal usability evaluation of their described EUD-ESM tools [101, 153].

Given the lack of evaluation research, it is unsurprising that only the work of Ludwig et al. [101] could be considered *Describing* research by Kjeldskov and Graham. A publication with the purpose of describing ideal system properties requires supporting evidence of what properties are actually beneficial, based on empirical evaluation. Although the surveyed publications often state that informal usability testing has been conducted with researchers, these claims do not advance the field in terms of best practice for how to model an ESM study in such a way as to be understood, and thus created and modified, by non-programmers.

The concept of usability has been defined in a variety of ways, often in terms of the ISO standard, which states that usability is *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* [175, p. 537]. Usability is thus largely context-dependent, and also depends on the measures used to assess this effectiveness, efficiency and satisfaction. Hornbæk provides a comprehensive review of the methods used to measure these three constructs [176]. Different research methods were also undertaken within the studies in this chapter. In the first study, quantitative measures such as task time and error rate were taken. Conversely, the final study did not quantify any such measures for statistical analyses, but instead took a qualitative approach by obtaining insight from participants' think-aloud during task completion and retrospective semi-structured interviews.

Table 6.1: The three usability studies of this chapter in relation to the usability constructs employed

	Study #1	Study #2	Study #3
Effectiveness	Task completion Accuracy of solution Errors made	Task completion Accuracy of solution	Task completion Accuracy of solution
Efficiency	Task completion time		
Satisfaction	System Usability Scale Post-study interview	System Usability Scale	Post-study interview

The three studies in this chapter are as follows: a study with 20 participants where quantitative metrics on task time and error rate were collected (Study #1, Section 6.2) a smaller study with six health psychology students (Study #2, Section 6.3) and a final study with 10 participants that collected more qualitative data (Study #3, Section 6.4). The studies are summarised in Table 6.1 with respect to the research methods, relevant to each of the three usability constructs, that they employ.

6.1.1 How important is usability?

As discussed in Section 4.1, evaluations of EUD tools are most frequently conducted within the constraints of lab-based usability studies. However, such studies are limited in their capacity to determine the acceptance of an EUD tool in its intended context of use. As highlighted in [143], many products are commercial successes despite being difficult to use, and conversely, highly *usable* products may fail if they are not *useful* in practice.

Nevertheless, in the case of an EUD tool, usability studies are necessary. Thoroughly researched models of technology acceptance all consider ease-of-use as a key prerequisite for the adoption of a new technology, particularly in its early stages. For example, "perceived ease-of-use" is a core construct of the Technology Acceptance Model [145]. The Unified Theory of Acceptance and Use of Technology has the similar construct of "effort expectancy" [146]. The purpose of EUD is to make *useful* software functionality accessible to its end-users. Thus, if usability issues are not addressed in an EUD tool, then these potentially useful features are still inaccessible to end-users, who are likely to abandon unusable technology in favour of delegating development to a professional.

This chapter thus describes three formal usability studies of Jeeves conducted during its iterative development. In doing so, insights into the benefits and challenges of the blocks-based programming paradigm were obtained, as well as preliminary feedback on the real-world utility of such a tool. In answering the research sub-question "*How can the development effort* *be reduced to allow rapid creation of experience sampling apps with no programming?*" these studies aim to evaluate the implementation decisions discussed in the previous chapter.

6.2 Study #1: Feasibility and Quantitative Measures

The first iteration of Jeeves was evaluated in April 2015. As the first usability study of the prototype visual programming environment, its primary aim was to assess the general usability and learnability of the blocks-based paradigm. A number of introductory computer science courses introduce blocks-based environments to students prior to textual programming, highlighting their accessibility to non-programmers [138]. The popularity of *Scratch* is also testament to the the usability of blocks by children of all ages. However, it was not clear whether these properties would hold in the specific domain of ESM app creation. For example, studies supporting the usability of blocks environments are often in the context of a formal educational program, where such environments are employed over a number of weeks. Furthermore, the colourful, animated applications, and peer feedback employed in *Scratch* are motivational incentives for young users to engage with EUD activities. In the absence of these incentives, it was unclear whether a blocks-based tool would still be a viable method of ESM app creation.

The objectives of this study were to determine whether participants could complete the study tasks successfully, which were designed to represent EUD activities in a real-world situation, and to identify salient usability issues encountered in doing so. In obtaining quantitative usability measures, a secondary purpose of the study was to make an empirical comparison of effectiveness, efficiency and satisfaction between participants with programming experience, and those without. Thus, three hypotheses were tested by collecting quantitative data on task time, number of errors made, and usability as determined by the System Usability Scale [177]:

- **H1:** Non-programmers will have a significantly lower usability score than programmers.
- **H2:** Non-programmers will make significantly more task errors than programmers.
- **H3:** Non-programmers will take significantly more task time than programmers.

6.2.1 Participants

20 students from the University of St Andrews were recruited by convenience sampling. Their ages ranged from 19 to 34, and they had mixed experience of programming. 11 had prior experience of any sort of programming, with self-assessed confidence varying from beginner to advanced. All 11 programming participants had at least basic experience with imperative languages such as Java, Python and MATLAB, and the remaining nine had no prior programming experience. Participants studied a wide range of subjects, including computer science, physics,

chemistry, history, English, management and international relations. To ensure consistency of experience, none of the participants were familiar with ESM.

6.2.2 Data collection and analysis

This was a within-subjects study, with each participant exposed to the same version of Jeeves, and asked to complete the tasks in the same order. However, for the purposes of analysis, participants with programming experience were compared to those without. Although it has been observed that testing with five participants is enough to uncover 80% of usability issues [178, 179], the analysis of this study involved making statistical comparisons between the 9 non-programmers and 11 programmers, for which larger sample sizes would be desirable for reducing errors. Despite this limitation, the primary purposes of this study was to identify *practically* significant differences between the two groups, as well as salient usability issues.

It was considered necessary to test the usability of Jeeves with participants from a variety of backgrounds, and varying experience of programming. Although the participants of this study were not the intended users of Jeeves (namely psychology researchers and clinicians) domain knowledge was not necessary for the purposes of assessing usability of the blocks-based paradigm. As such, this study provided participants with specific tasks that did not require an understanding of ESM or its applications.

Study tasks were completed in a room where only the researcher was present, on a machine running screen capture software to accurately collect participants' task times and instances of errors. Audio capture was also used to record post-study interviews for later transcription. With this procedure, there are potential sources of bias. First, while participants were instructed to complete tasks in their own time, it is possible, particularly given the researcher's presence, that they felt pressured to complete tasks quickly. Similarly, it is possible that there was response bias in post-study feedback, whereby participants adapted their responses to be more favourable towards Jeeves. Finally, as only the researcher viewed the screen captures and categorised the types of errors made, the results lack inter-rater reliability.

6.2.3 Description

Participants first completed an introductory questionnaire in order to obtain demographic information, prior programming experience, and general use of technology. Participants also reported whether they had used reminder apps, self-monitoring apps, or automation apps before, as it was thought this may affect participants' understanding of the real-world utility of Jeeves.

Participants were then given an introduction booklet, which provided a basic explanation on

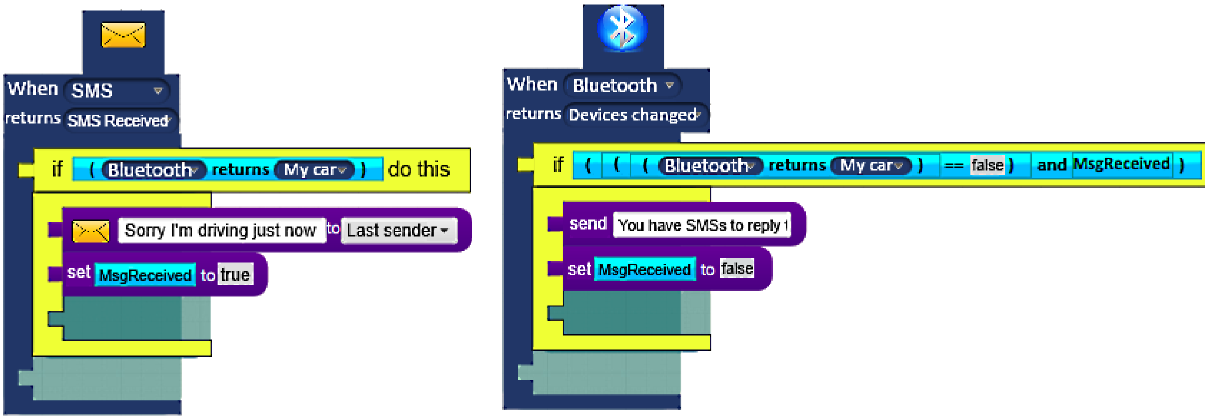


Figure 6.1: Specification presented to participants in Task 2

how to use Jeeves, with respect to dragging and dropping blocks together, and creating surveys. In this booklet, Jeeves was introduced as a tool to create apps for self-monitoring, in order to provide participants with a better understanding of the usefulness of Jeeves. Participants were given 10 minutes to familiarise themselves with the information in this booklet. Following this, they were then issued a study script detailing four tasks, and were instructed to work through these tasks at their own pace. A maximum of 30 minutes was allotted to this part of the study to allow time for post-study interviews.

6.2.3.1 Task 1 - Step-by-step guide

The purpose of this task was to familiarise participants with the layout of Jeeves, while isolating issues associated with the control and navigation, thus no issues in language comprehension were highlighted. Instead, the task was designed to cover all testable features of the tool.

Participants were given a separate booklet that guided them through the creation of a basic application, with specific instructions on where to find relevant blocks and how to connect them together. This task was also intended to provide participants with an introduction to the self-monitoring capacity of Jeeves apps, so that an indication of participants' understanding of the utility of Jeeves could be obtained in post-study interviews.

6.2.3.2 Task 2 - Reading and explanation

This task asked participants to read a previously created study specification, and to provide a textual description of their interpretation of its function. The specification used was designed to send an automated reply to any SMS sent to the user while driving. (In this case, the user is assumed to be driving when their device's Bluetooth sensor detects another Bluetooth-enabled device in their car.) The specification presented to participants is shown in Figure 6.1, which also illustrates some visual design changes that have been made to Jeeves since this study was conducted.

In addition to making applications easy to design and develop, it was also considered important to ensure readability of the created specifications, to support collaboration between researchers. While *writability* and *readability* design decisions were discussed in the previous chapter, this task sought to validate the readability decisions made with respect to the Physics of Notations [162].

6.2.3.3 Task 3 - Existing study modification

In this section of the study, participants were asked to augment the specification of Task 2 with additional features. Again, with a view towards allowing collaboration between developers, it is important that users can create their own specifications, but also provide useful edits and contributions to *existing* specifications. The functionality requested consisted of: turning on the device's speakerphone on beginning a call, creating a new survey, and sending this survey when the user leaves their car.

6.2.3.4 Task 4 - Freeform design task

This task asked participants to design and create an application with no prior structure. Their application was to be based on a paragraph describing a use-case scenario, involving a woman called Susan and her elderly mother, Ethel. The required specification was described to participants as follows:

"Design an app that reminds Ethel to take her medication in the morning, afternoon and evening. The app should also ask her every few hours if she's okay. If she doesn't respond after a certain time, the app should let Susan know that Ethel has been unable to respond"

Although this specification was suggestive of required blocks, no concrete implementation structure was given and participants were free to use their creativity to provide a solution.

6.2.4 Quantitative results

Both quantitative and qualitative results were collected in this study. The quantitative data collected were on task time, error rate and usability, as evaluated by the System Usability Scale. In addition, post-study feedback was acquired from unstructured interviews.

Charts illustrating the comparative results of programmers and non-programmers for error rate, task time and SUS score are shown in Figures 6.2, 6.3 and 6.4 respectively. Error rates and task

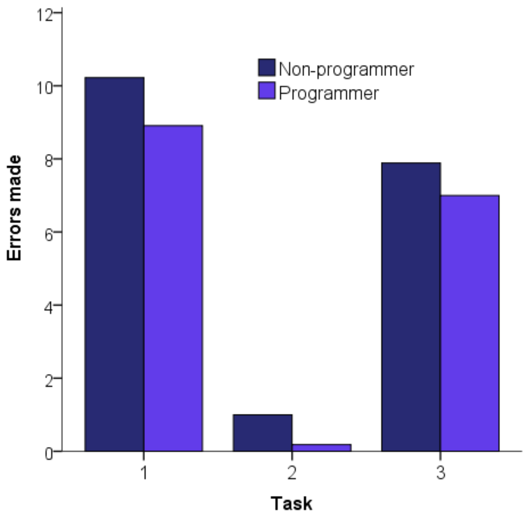


Figure 6.2: Errors made in the first three tasks

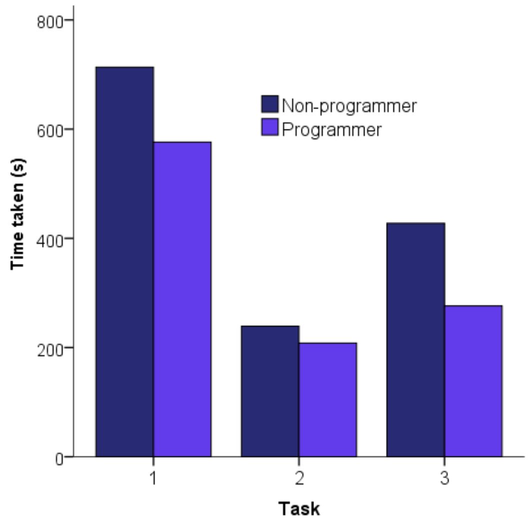


Figure 6.3: Time to complete first three tasks

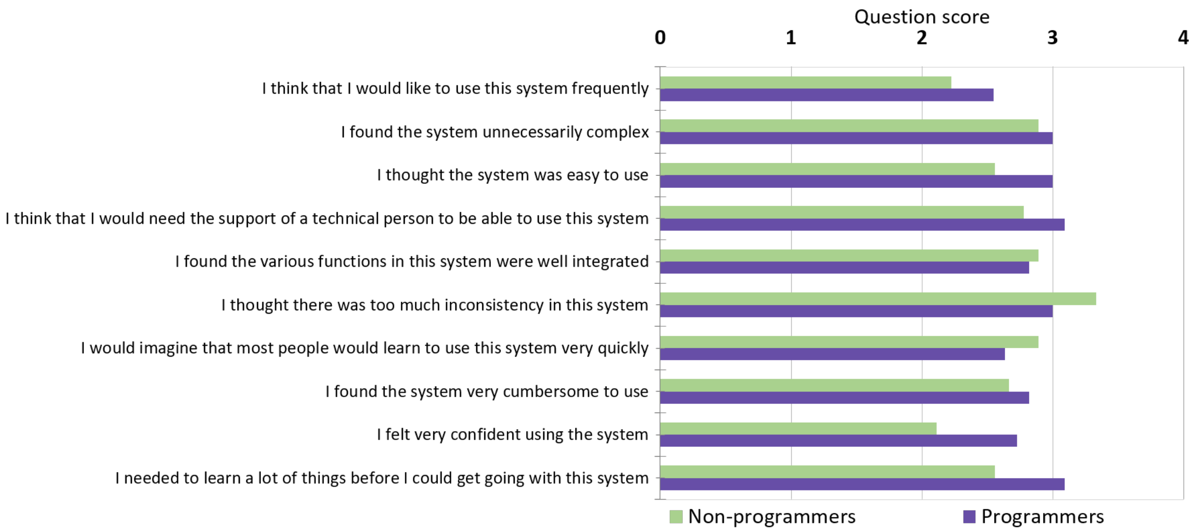


Figure 6.4: SUS questions with scores from programmers and non-programmers

times were not counted for the final task, as not all participants were able to complete this task in the allotted time. A summary of quantitative data and statistical test results is given in Table 6.2.

6.2.4.1 Errors made

In general terms, errors made by participants were instances where they performed an action that hindered progress to successful task completion. Through a process of coding errors observed in the screen capture recordings, three types were identified. *Presentation errors* were all those that were caused by the presentation of information to the participant, including instances where a participant misinterpreted a button's function or omitted a step in completing a task. *Navigation* *errors* were all attempts to find a block in the wrong menu, or a navigation to the incorrect sidebar tab. (A sequence of incorrect menu activations was counted as one error.) *Control errors* were those made in direct manipulation of blocks, such as dragging immovable buttons, or accidentally removing blocks from their nested location. Importantly, there were no significant differences in the number of each error type made between programmers and non-programmers.

For Tasks 1, 2 and 3, programmers made an average of 8.9, 0.2 and 7.0 total errors. Non-programmers made an average of 10.2, 1.0 and 7.9 errors respectively. Independent samples t-tests were run for each of the three tasks, showing no significant differences for Task 1 ($t(18) = 0.72, p = 0.481$), Task 2 ($t(18) = 1.66, p = 0.114$) or Task 3($t(18) = 0.52, p = 0.608$).Thus, the second hypothesis (*Non-programmers will make significantly more task errors than programmers*) is **rejected**. Results are illustrated in Figure 6.2.

6.2.4.2 Task Time

For Tasks 1, 2 and 3, programmers took an average of 576.2s, 208.1s and 276.3s respectively, while non-programmers took an average of 713.4s, 239.2s and 427.4s respectively. As before, three separate independent samples t-tests were run. In this case, the tests showed no significant difference in time taken for Task 1 ($t(18) = 1.89, p = 0.075$) or Task 2 ($t(18) = 1.35, p = 0.194$), but a significant difference for Task 3 ($t(18) = 2.21, p = 0.040$), such that the third hypothesis (*Non-programmers will take significantly more task time than programmers*) **cannot be rejected**. A chart of these results is shown in Figure 6.3.

6.2.4.3 Usability

The System Usability Scale (SUS) is referred to as a “quick and dirty” means of evaluating the usability of a system, and consists of a simple 10-item Likert scale questionnaire, which has been shown to be a robust and reliable measure of usability [177]. Each question has a numeric response option from 1 (strongly disagree) to 5 (strongly agree). For each of the negatively phrased questions (those with even numbers) this score is reversed, such that a high score on each question is desirable. In this study, programmers and non-programmers reported numerically similar scores ($M = 71.8, SD = 10.7$ for programmers, and $M = 67.2, SD = 13.9$ for non-programmers). To compare the two groups, a Shapiro-Wilk test was first run to test for normality of score distribution. As the null hypothesis of a normal distribution cannot be rejected in the programmers’ scores ($p = 0.577$) or the non-programmers’ scores ($p = 0.111$), an independent samples t-test was used. This t-test showed that there was no significant difference in usability between the two groups, $t(18) = 0.838, p = 0.413$, so the first hypothesis (*Non-programmers will have a significantly lower usability score than programmers*) is **rejected**.

Table 6.2: Results of t-tests on quantitative data (significant difference highlighted in red)

	Programmers	Non-programmers	Result of t-test
Task 1 time (s)	576.2 (128.6)	713.4 (195.3)	p=0.075
Task 2 time (s)	208.1 (31.3)	239.2 (68.5)	p=0.194
Task 3 time (s)	276.3 (63.4)	427.4 (216.6)	p=0.040
Task 1 errors	8.9 (3.6)	10.2 (4.6)	p=0.481
Task 2 errors	0.2 (0.4)	1.0 (1.6)	p=0.114
Task 3 errors	7.0 (2.4)	7.9 (5.0)	p=0.608
SUS score	71.8 (10.7)	67.2 (13.9)	p=0.413

A previous survey computed the mean SUS score across over 2000 studies to be 69.7 [180], and an acceptable usability score is defined to be 70 or above. Hence, these results suggested that at this early stage of development, Jeeves had acceptable usability to those with programming experience, but still required work to be acceptable for non-programmers, for whom usability is of high importance. Results for both programmers and non-programmers on each question are shown in Figure 6.4.

Further, research by Lewis and Sauro [181], has shown that the SUS can also be used to separately evaluate "learnability", or how easily the system can be learnt. Learnability is of particular relevance to the work in this thesis, given the time constraints of researchers and clinicians, and an inclination to delegate such development to professional programmers or commercial tools. Thus, minimising the time in which researchers could learn to use Jeeves was assumed to be critical to its adoption. Learnability scores were computed as a score out of 100 based on answers to Questions 4 and 10 of the SUS, as described by Lewis and Sauro. The mean learnability was 77.3 ($SD = 13.5$) for programmers and 66.7 ($SD = 21.7$) for non-programmers, which again suggested that Jeeves was learnable by programmers, but required additional design modifications for non-programmers.

6.2.5 Qualitative results

These quantitative results were indicative of an acceptable level of usability and learnability for programmers, but still insufficient for non-programmers. In order to identify subjective perceptions of problems and how they could be resolved, qualitative results were obtained from triangulating observation of participants completing the tasks with their post-study interview feedback. In general, given that the first three study tasks were relatively instructive with regards to the blocks required, the majority of barriers encountered were related to the direct manipulation

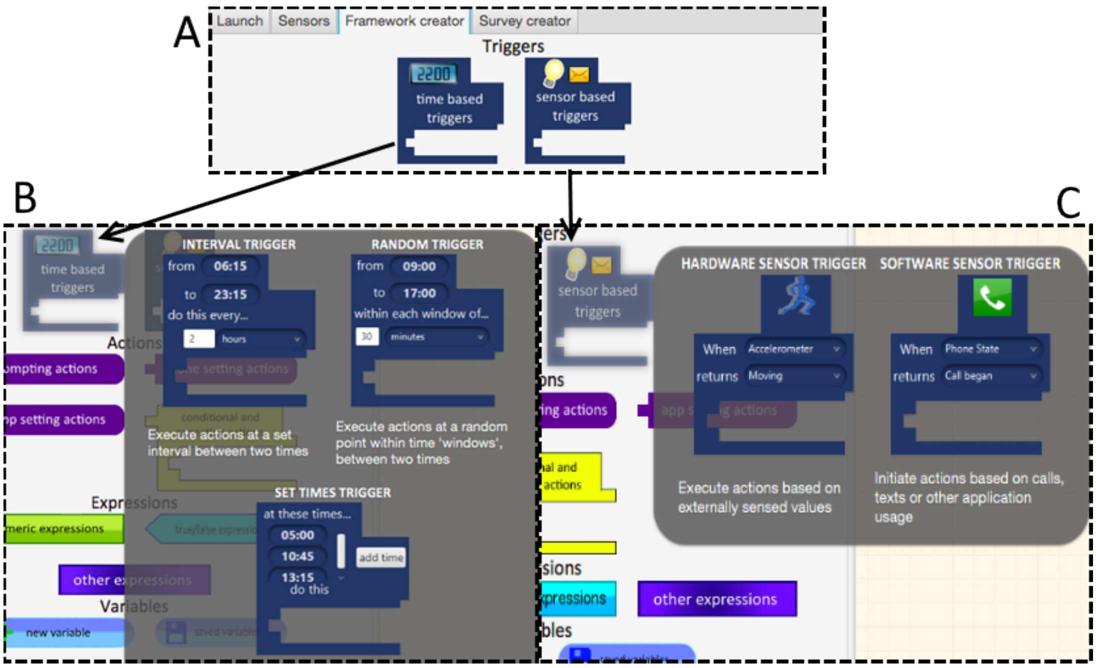


Figure 6.5: Menu system in Study #1, where different components are obscured in menus

of the interface, and not to participants' problem solving capabilities with Jeeves. Salient issues noted are described in detail as follows.

6.2.5.1 Issue 1: supporting exploration

The exploratory nature of non-programmers appeared to be disrupted by visibility issues, an example of which is shown in Figure 6.5, where blocks are accessed from hidden menus. Figure 6.5A shows the initial menu view, consisting of two "abstract" triggers - namely time-based and sensor-based triggers. By clicking on these, their underlying "concrete" triggers are exposed, shown in Figure 6.5B and C respectively. For example, "Interval", "Random" and "Set Times" triggers are accessible from the time-based abstract trigger. Similarly, concrete actions and expressions were also grouped into menus, initially hidden until their abstract parent was clicked. Participants who had enough time to generate a partial solution to the final task tended to employ a "bricolage" approach to programming, as described by Turkle and Papert [164]. In this approach, participants build spontaneously from the ground up, deleting erroneous components as necessary, similar to participants in an analysis of Scratch user habits by Meerbaum-Salant et al [182]. Notable exceptions were the three participants who considered themselves to be advanced programmers, who appeared to have a pre-conceived design, and constructed their solutions in a linear process.

The previous palette design, as shown in Figure 6.5, had blocks organised into menus, which were made visible by clicking on their abstract parent block, such as those labelled "time-based triggers" or "prompting actions". Although this menu structure was intended to reduce the time needed to find relevant blocks, it appeared to have the opposite effect, due to participants' tendency to search through menus at random. This difficulty was expressed by participants in their post-study feedback:

"...everything is very intuitive the building but the categories? It's hard, I dunno. But maybe it's just me" (P18, programmer)

"I was slightly confused by the classifications of things like app setting actions...just like 'set' together with 'wait for' and the same time there's like the 'sensor action' things" (P14, programmer)

Quantitative results showed that programmers were significantly faster than non-programmers, but with a similar number of errors. Before reviewing the screen recordings again, it was hypothesised that this was due to programmers having a greater recall of required components (each sequence of incorrect menu navigations was counted as one *navigation error*, so this suggested that programmers' sequences of wrong menus were shorter). Recordings were reviewed and the number of incorrect navigations – entering a menu then exiting immediately – were counted. This figure was divided this by the time taken to give a rate of incorrect clicks. Surprisingly, programmers' average rate of incorrect clicks was 1.06/minute, and non-programmers' average rate was 0.95/minute, with no statistically significant difference between the two. From observation it was notable that non-programmers would often pause before deciding whether to click on a menu button, whereas programmers were inclined to explore without hesitation.

This issue suggested that menus were unnecessary, particularly if superfluous blocks were removed. For example, Moody's Principle of Perceptual Discriminability: *"Different Symbols Should Be Clearly Distinguishable from Each Other"* is supported, because the text and widgets incorporated into each trigger and action are used as distinguishing features. However, as shown in Figure 6.5, "Hardware Sensor Triggers" and "Software Sensor Triggers" are almost identical, and participants frequently used the wrong type, which aggravated confusion.

Resolution - In order to resolve this issue, the two sensor-based triggers were incorporated into one "Sensor Trigger" block. Further, the abstract block buttons were removed altogether and replaced with palettes from which blocks could be dragged directly, as shown in Figure 6.6. For example, all actions can be accessed from the palette labelled "Actions". As such, the Cognitive Dimension of *visibility* was expected to be improved.

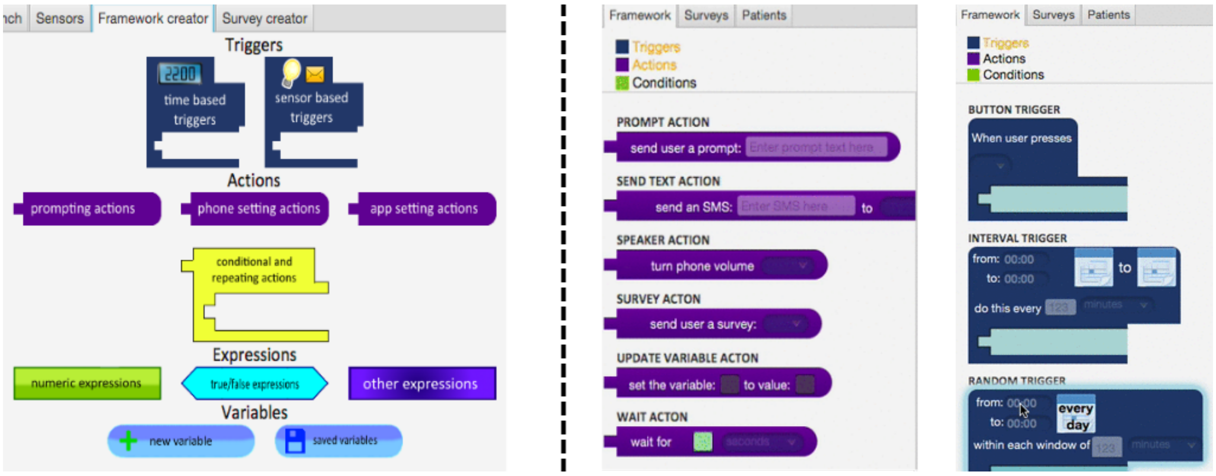


Figure 6.6: The single menu of abstract blocks was replaced with separate labelled menus containing trigger/action/condition blocks

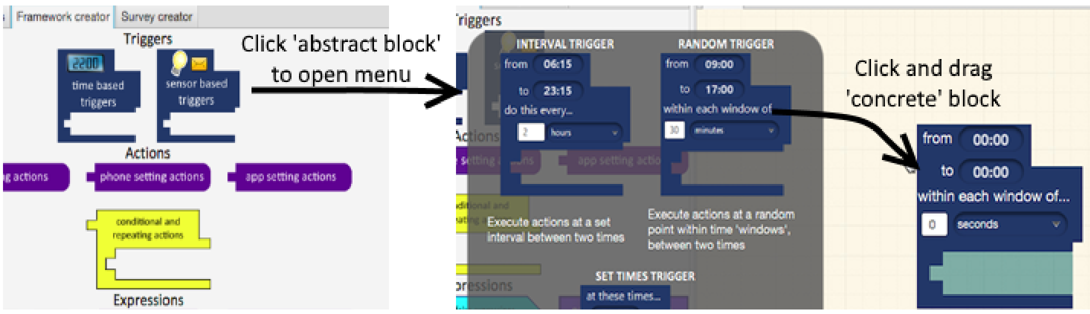


Figure 6.7: The previous click-click-drag process caused issues for participants

6.2.5.2 Issue 2 - Improving Control

The most prominent issue that participants had was with direct manipulation of the blocks themselves. The majority of participants explicitly mentioned having difficulties with the drag-and-drop process, and all were observed to experience issues, again aggravated by the menu design. In this design, participants had to click the 'abstract block' to open a menu, and then click and drag the 'concrete block' from the menu onto the canvas. This sequence, as illustrated in Figure 6.7, was a notable issue for both programmers and non-programmers alike:

"sometimes you'd be trying to click something and you'd actually move something and then...just little errors might be magnified by the fact that it's very click-based" (P17, non-programmer)

"...you need to click and then drag. I always try to just click and click and then I'm expecting something to happen...because you're kind of primed oh you click and then you click and then, oh then actually you have to drag" (P16, programmer)

The abstract blocks, labelled with concepts such as "time based triggers", did not respond to

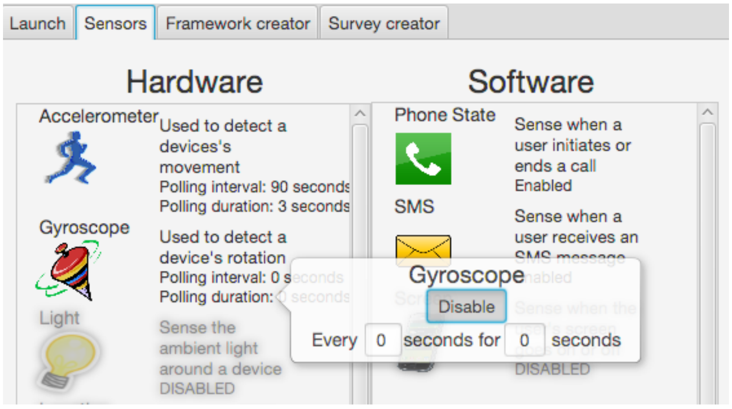


Figure 6.8: The sensor configuration tab, which was removed in later iterations of Jeeves

mouse input in the same way as the concrete blocks, and instead acted as buttons for accessing menus of these concrete blocks. This representation was intended to support the Cognitive Dimension of “abstraction”. Further, designing these buttons to resemble the concrete blocks contained in their respective menus was intended to support *role-expressiveness* while also adhering to Moody’s Principle of Semantic Transparency: *“Use Visual Representations Whose Appearance Suggests Their Meaning”*. For example, the appearance of the “time based triggers” button was intended to suggest its role as a means to display different concrete time triggers. However, this design decision had the inverse effect, as the appearance of the button suggested that its role was actually as a draggable block.

Resolution - As described in the resolution of Issue 1, abstract blocks were removed, such that all concrete blocks are now directly draggable from their respective palettes. It is also possible that this issue would be resolved by having menu buttons actually resemble buttons, rather than draggable blocks.

6.2.5.3 Issue 3 - Too Much Information

Multiple participants, particularly non-programmers, mentioned how they initially struggled with the number of new concepts. For example, one non-programmer suggested that: *“at first it’s maybe a wee bit, a lot to take in. The different vocabulary of it all, and what different things mean”* (P7, non-programmer). Those who did not struggle expected that others with less programming experience would face such issues: *“I just think that if I gave that to my mum or dad I don’t think they’d be able to start with it”* (P4, programmer).

Moody’s Principle of Semiotic Clarity, *“There Should Be a 1:1 Correspondence between Semantic Constructs and Graphical Symbols”*, as well as the *“closeness of mapping”* Cognitive Dimension, both influenced the original design, but apparently to a detrimental effect. In attempting to capture all semantic constructs, including loops, arithmetic expressions, and various smartphone system actions, participants expressed concern that there were too many concepts introduced at once.

From this feedback, the decision was taken to minimise the number of constructs that the end-user would have to comprehend, in line with Moody's design Principle of Graphic Economy: *"The Number of Different Graphical Symbols Should Be Cognitively Manageable"*. For example, the encoding of sensors was significantly simplified. Figure 6.8 shows the content of the sensor configuration tab previously accessible in Jeeves. The purpose of this pane was to allow end-users to enable and disable specific phone sensors, as well as the frequency and granularity of sampling. However, one participant highlighted this as a potential example of *"hidden dependencies"*:

"Only thing that confused me a little bit I was looking for like, sensors right? Okay this was something actually, to be fair I should have been prompted to enable this kind of stuff right?" (P18, programmer)

Resolution - The sensor configuration tab was removed, as well as constructs considered to be superfluous. For example, the constructs representing loops were removed, as these were observed to cause the most confusion for non-programmer participants, to whom these concepts had not previously been introduced. Furthermore, in designing tasks based on ESM studies, it was unclear whether loops would serve a useful purpose.

6.2.5.4 Positive Feedback - Learning curve

Regardless of task completion or salient difficulties in doing so, the majority of participants expressed that they felt Jeeves had a shallow learning curve, and non-programmers had consequent feelings of improvement towards the end of the study:

"...cause I'm not really a computer person, I think it was initially a bit overwhelming but once I read the instructions and familiarised myself it was really good" (P11, non-programmer)

"maths and like, concepts of y'know, if this then this happens and this, like box and sort of diagrams like that, are a bit confusing, but I felt that this got better a bit at the end" (P8, non-programmer)

Given that participants only had 30 minutes of time using Jeeves, such feedback was particularly positive. It is important that end-users with no prior programming experience are able to learn the basics of Jeeves quickly and efficiently, if it is to be adopted in practice. When asked how the learnability of Jeeves could be improved, participants were uncertain, and suggested different features, including built-in tutorial features, and different means of selecting blocks.

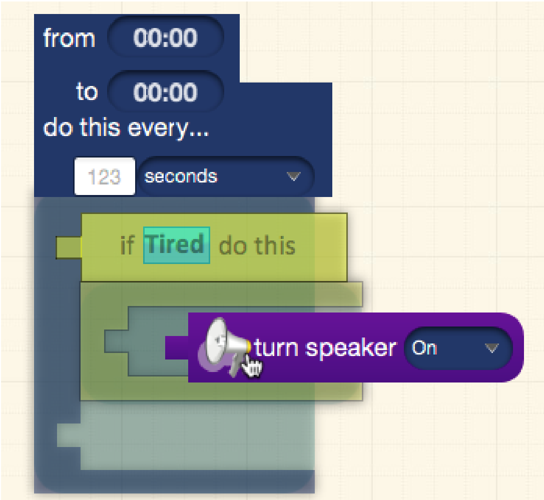


Figure 6.9: Old drag-and-drop highlighting caused ambiguity in Study #1

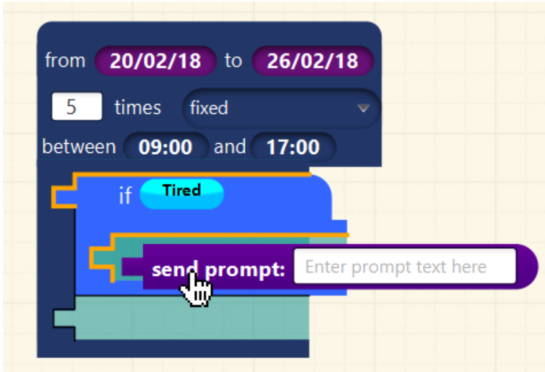


Figure 6.10: Updated highlighting improved clarity in later studies

6.2.5.5 Positive Feedback - Visual metaphor

The blocks-based paradigm received almost unanimously positive feedback, with comments made by programmers and non-programmers alike on the intuitive means of fitting blocks together. Indeed, only one participant, an experienced programmer, expressed dislike of visual programming, explaining: *"I don't like having to click and then click and drag. It doesn't work for me...I'm a programmer. I'd rather just write a loop myself"* (P2, programmer).

P7, with no programming experience, appreciated the visibility and 'flow' of the paradigm: *"...rather than just simply words and numbers and things like that you feel like you can sort of see it and see the flow of ideas and things and the sort of different steps"* (P7, non-programmer).

Having confirmed the feasibility of the blocks-based notation through this encouraging feedback, very little action was taken with regards to the paradigm itself. However, it was observed that participants frequently struggled to drag and drop action blocks into the triggers or nested if-conditions. Figure 6.9 shows the previous visual feedback when an action block is dragged over an action receiver, which was observed to cause confusion. To rectify this, a clearer notation was used to highlight the relevant block, shown in Figure 6.10. This was expected to reduce the Cognitive Dimension of *"error-proneness"* of the previous approach, by clarifying where a dragged action would be positioned, an issue expressed by one participant:

"Sometimes you have to get this thing right in the right place or else it doesn't, like if you kind of like drag it and drop it like here it won't register" (P12, non-programmer)

6.2.6 Discussion

The results of Study #1 demonstrated the potential of Jeeves as an environment usable by those with no prior programming experience. As a feasibility study for developing Jeeves further, it confirmed that the blocks-based programming paradigm was approachable by non-programmers. Through both direct observation of participants' usage of Jeeves, and from analysis of their post-study feedback, actionable results were obtained that directed the second iteration of Jeeves.

An interesting finding of this study was that design decisions guided by the Physics of Notations and Cognitive Dimensions were not always well-received. This was due partly to the acknowledged trade-offs in dimensions; for example, the Principle of Semiotic Clarity conflicts with that of Graphic Economy. Similarly, the *abstraction* dimension conflicts with the *visibility* dimension. (See Section 5.3 for a discussion of guideline trade-offs.) However, design principles related to semantic transparency and role-expressiveness were found to be self-conflicting in Jeeves. For example, while the appearance of the "time based triggers" button implied its role for showing concrete time triggers, it also *incorrectly* implied its role as a concrete block itself.

6.2.6.1 Limitations

A key limitation of this study (and indeed all three of the usability studies in this chapter) is that participants were students rather than clinicians and psychology researchers, which reduces the study's external validity. The primary reason for doing so was that students were easily accessible. Given that this was an initial proof-of-concept and usability study, the ability to acquire larger numbers of participants from more diverse subject backgrounds and exposure to programming was considered to be of high importance. It allowed a variety of feedback to be collected, ranging from those with no programming experience and limited computer use, to HCI students who had studied interface usability as part of their undergraduate or postgraduate work. As such, a larger number of usability issues were identified which, if unresolved, could significantly reduce the utility of Jeeves as perceived by its target end-users. However, as previously discussed, this was a small number of participants for the purposes of testing for statistical significance.

It could also be argued that the students had no knowledge of experience sampling, so that by asking these students to create experience sampling apps, this would not represent the real-world use of Jeeves. To mitigate this, Jeeves was introduced to study participants as an application to create self-monitoring and automation apps. Only one participant explicitly stated that she would not use it in her daily life, and indeed four explicitly stated, without being asked, that they would like to use Jeeves for creating their own apps. Given the usability issues encountered, the enthusiasm for real-world usage by participants was surprising, and suggested an application of

Jeeves as a tool for automating smartphone functions, similar to popular apps such as *Atooma* and *Tasker* [183]. While such an application of Jeeves was not pursued in this thesis, it is considered to be potential future work.

6.3 Study #2: Health Psychology Student Workshop

The second usability evaluation took place in November 2016, prior to which the usability issues salient in Study #1 had been resolved, and some changes to visual appearance implemented. Again, although researchers and clinicians would have been preferable participants in this study, it was decided that further issues would likely be discovered, that would require resolution before such an evaluation would be useful.

The procedure for this evaluation was adapted for the context in which it took place. Rather than testing with each participant individually in separate study sessions, this evaluation was conducted in a workshop format, whereby all participants were in the same room, and completed the study tasks simultaneously over a two hour period.

6.3.1 Participants

Participants were six medical postgraduates, studying for a Masters in Health Psychology. As part of their schedule, the students had weekly two-hour tutorials, one of which was used to conduct the Jeeves workshop. Of the six participants in the tutorial group, four had no prior programming experience, and two had limited experience in web programming. As in the previous study, none of the participants were familiar with experience sampling. However, as students pursuing a degree in Health Psychology, these participants were considered to be potential future users of Jeeves, unlike the previous study where participants were largely students of subjects unlikely to apply the method.

6.3.2 Workshop structure

The workshop began with participants completing a questionnaire on their computer usage and programming experience, as well as demographic information such as gender and spoken languages. A 15 minute presentation was then given, which defined and described experience sampling, and provided a short summary of Jeeves itself.

Following this presentation, participants were asked to download the latest version of Jeeves on their own laptops, and to access a provided URL that directed them to the first study task. These tasks were presented to participants in web pages, unlike the previous study, where tasks and guides were provided in paper form. It was observed that, as a static medium, the paper booklet did not clearly communicate how the blocks were to be dragged and dropped together, which potentially aggravated issues of direct manipulation. In presenting web-based instructions, it was possible to use GIF animations to demonstrate the required actions to participants. Participants were not given an explicit time limit, although following the presentation and installation of Jeeves, 90 minutes remained in the workshop for tasks and post-study feedback.

6.3.3 Data collection and analysis

Like the previous study, this was a within-subjects study, with each participant exposed to the same version of Jeeves, and asked to complete the tasks in the same order. Further, for the task where participants used *MovisensXS*, an alternative tool for ESM app creation, all completed this task having completed the previous two tasks with Jeeves. While this introduces bias, because the aim was not to make an empirical comparison between the two tools, any such bias was permissible. During the study period, participants did not talk to one another, and were only given assistance to download Jeeves onto their laptops at the beginning of the study.

The primary purpose of this study was to identify salient usability issues, therefore six participants was considered to be adequate for identifying a large number of such issues. However, more participants would have enabled identification of less common, but potentially serious problems with Jeeves.

Study tasks were completed in a room where the researcher was present, and all participants worked through the tasks in the same two hour session on their own laptops. Due to this, screen capture software was not installed, and instead post-study survey feedback, and direct observation of participants interacting with Jeeves, were the primary sources of data for analysis. Also, as each participant was assigned a unique log-in ID and password for Jeeves, it was possible to view their completed study specifications by logging back in with these credentials after the study's completion.

6.3.3.1 Task 1 - Walkthrough

As in the previous study, the first task guided the participants through the creation of a basic Jeeves ESM application, in order to fully introduce the testable features of the environment.

Novel testable features of this version of Jeeves were the Log Design panel and Button Triggers, as described in Section 5.4, allowing researchers to implement event-contingent sampling. Although students were not asked to create and set attributes, in this iteration of Jeeves it was possible to add numeric survey answers to a pre-created "survey score" attribute. This functionality was inspired by interventions whereby a patient who scores highly on a validated measure of mental or physical distress is automatically provided with resources to help them.

6.3.3.2 Task 2 - Study replication

In this task, five publications that described ESM studies, taken from medicine and psychology journals, were provided to the participants [184, 185, 186, 187, 188]. The information relevant to the ESM study specification was highlighted in each publication so that participants did not waste unnecessary time searching for it.

Participants were asked to create a specification in Jeeves, based on descriptions in each of the five studies, in order to assess whether participants could apply Jeeves for its intended purpose. The studies described in these publications incorporate signal, interval and event-contingent triggers, as well as surveys and reminder prompts. For example, the following paragraph was highlighted in the study by Wonderlich et al [188]:

"...all three studies included three types of daily self-report methods: signal-contingent (i.e., ratings were provided in response to six semi-random signals throughout the waking hours of the day), event-contingent (i.e., participants were asked to record the occurrence of binge eating and certain other eating disorder behaviors), and interval-contingent (i.e., participants were asked to complete EMA ratings at the end of each day)" [188, p. 309]

Given that many studies used standardised, lengthy questionnaires, participants were instructed to create one or two sensible questions for each survey required.

6.3.3.3 Task 3 - Interface comparison

Task 3 asked participants to use the *MovisensXS* online application discussed in Section 4.2. *MovisensXS* is a commercial ESM app creation tool that uses a diagrammatic visual programming language for constructing specifications, and is the only other example of a visual programming environment for such applications that is publicly available. Both its availability for trial use, and its visual programming paradigm, were key motivators for its choice. Further, MovisensXS has been employed in recent ESM studies surveyed in Chapter 2, demonstrating its real-world applicability (for examples, see [65, 189, 190]. Participants were asked to replicate the study described in the previous task by Wonderlich et al [188]. Example specifications of this study in *MovisensXS* and Jeeves are shown in Figures 6.11 and 6.12 respectively.

The purpose of this task was not to make a formal, empirical comparison between Jeeves and *MovisensXS*, but simply to get insight into the intuitiveness of the diagrammatic programming paradigm. Considerable additional work would have been required in order to make an

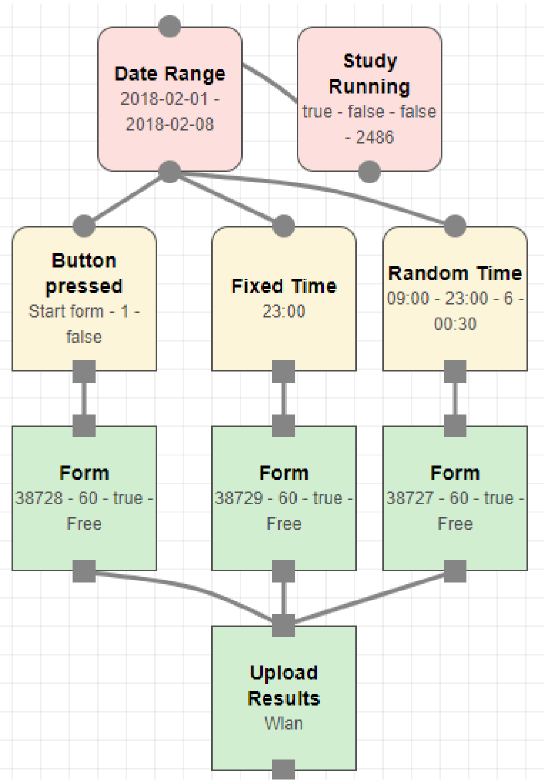


Figure 6.11: Study from [188] in *MovisensXS*

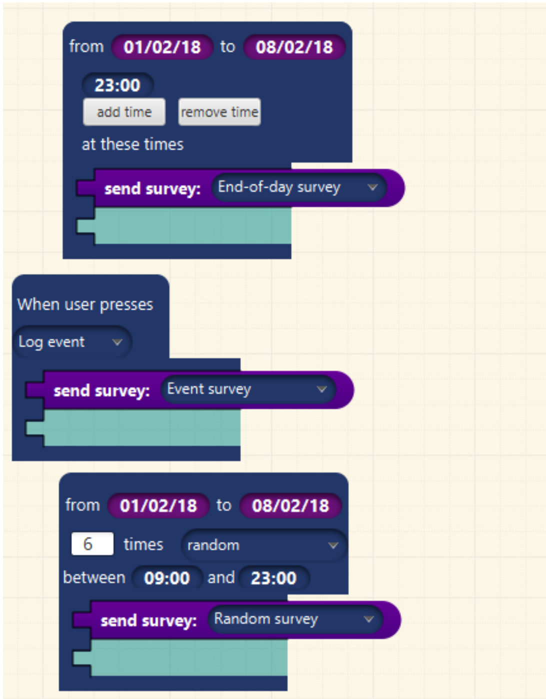


Figure 6.12: Study from [188] in Jeeves

empirically valid comparison, including a balanced introduction to both applications, and minimising learning effect, which was not possible given the time constraints of the study.

6.3.3.4 Task 4 - Free-form design task

Finally, participants were asked, based on their newly acquired knowledge of ESM and experience with Jeeves, to design and implement a study of their own conception. The purpose of this task was to assess participants' understanding of Jeeves as a potential research tool, given that they had practice of replicating previous studies. Unlike the free-form design task in Study #1, no use-case scenario was provided, as it was expected that participants could conceive their own, given their health psychology background.

Finally, a post-study questionnaire was completed online by the participants, which included questions of the System Usability Scale as applied in Study #1. As previously explained, these questions are phrased as statements, with response options ranging from 1 (strongly disagree) to 5 (strongly agree). Scores for negatively phrased statements are inverted so that a high score for all questions is desirable. Additionally, targeted questions addressing specific features of Jeeves were given, along with space for open comments and feedback.

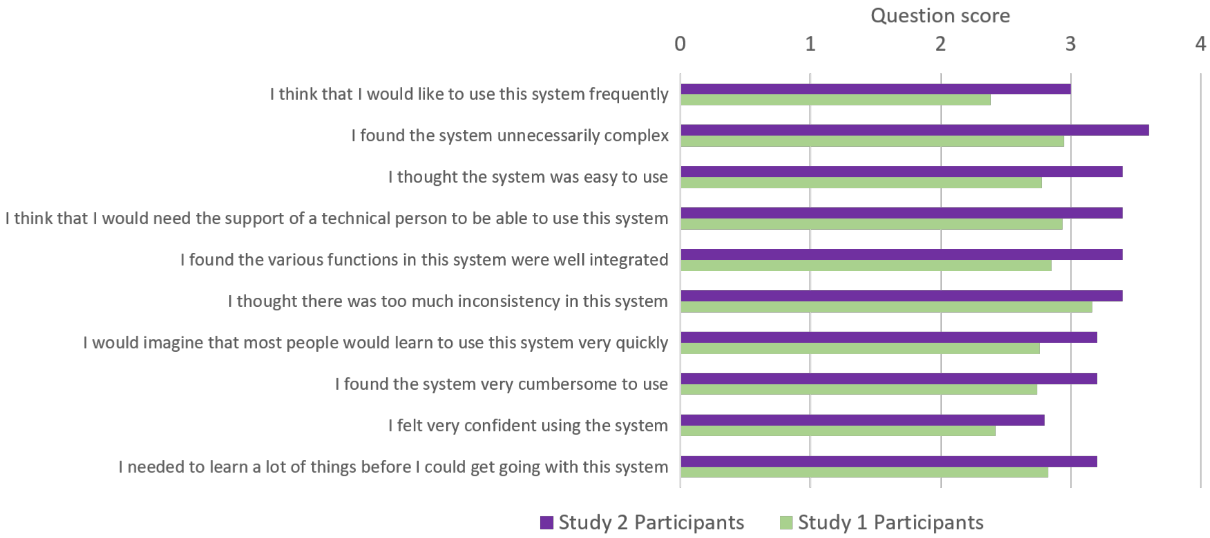


Figure 6.13: SUS question scores comparing participants in Study #1 and #2

6.3.4 Quantitative results

The results of the workshop are based both on the questionnaire responses from participants and assessment of their implemented specifications. As all six participants were working on the tasks simultaneously, and no screen capture software was utilised, it was difficult to observe the issues that participants encountered in their process of implementation. As such, the results of this study are less detailed than those of Study #1, where participants were observed, recorded and interviewed individually.

6.3.4.1 Satisfaction

In analysing participants' created specifications and written feedback, it was noted that one participant was a significant outlier. English was his third language and it was clear from issues he encountered during the study, and observation of his progress and results, that he struggled with the basic reading and writing skills that were required. His SUS score of 42.5 reflects the struggles he had and highlights that, although Jeeves is a visual language, its ease-of-use relies on end-users having adequate verbal skills.

The average SUS score for the participants was 75.7, a marked increase from that of 69.7 reported by participants in the previous study. Indeed, removing the aforementioned outlier, the average SUS score was 81.5, suggesting that the more prevalent usability issues highlighted in the previous study had been resolved. A chart comparing the score for each SUS question between participants in the two studies (with the outlier removed) is displayed in Figure 6.13.

After completing the SUS questions, participants answered similar Likert scale questions targeting specific functionality of Jeeves. The questions (with participants' average score out of a maximum of 4) are as follows:

- I found the Jeeves survey creation pane easy to understand (3.20 / 4.00)
- I dislike the appearance of the visual components (3.00 / 4.00)
- I think the block-based layout is intuitive (3.00 / 4.00)
- I don't think I could create basic applications on my own (2.60 / 4.00)

As with the SUS, questions alternated between positive and negative phrasing. In the two negative cases, the score polarity was reversed, thus it was observed that a generally positive response was given to all four questions.

It was important to question the intuitiveness of the survey creation pane separately, as it was not related to use of the blocks. However, it is notable that the clarity of the survey pane was rated particularly highly in its relevant question. A slightly lower rating was given on participants' confidence to create applications without guidance. This is unusual given that participants' score on Question 4 of the SUS – *"I think that I would need the support of a technical person to be able to use Jeeves"* – suggested that they were confident in using Jeeves independently.

Other quantitative measures such as task completion time or error rate were not taken in this study. Further, the strength of claims based on this study's SUS score is low given that only five participants were assessed, and the tasks were different to those used in the previous study. Nevertheless, the high usability scores, as well as observation of participants completing the tasks, suggest that the control and navigation issues that were prevalent in the previous iteration of Jeeves had been resolved.

6.3.5 Qualitative results

Qualitative results in this study were also less detailed than in the previous study. Without explicitly asking participants about their specific issues in an interview format as before, the received written feedback was considerably less detailed. Two general, open-ended questions were given to which participants provided written feedback - the first asked participants to compare Jeeves with *MovisensXS*, and the second asked for general feedback on Jeeves, and the workshop as a whole. Participants' feedback, although lacking detail, was largely positive, with the exception of the outlier, who skipped the first question and simply wrote *"need more training in order to use that properly"* in the second. As in the results of Study #1, salient issues are further described, with the steps taken to resolve them.

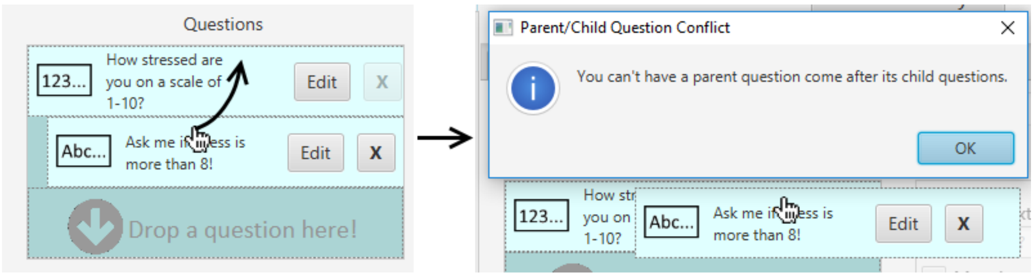


Figure 6.14: Reordering constraints were added to conditional questions

6.3.5.1 Issue 1 - Viscosity

It was promising that no participants explicitly mentioned control or navigation issues with the blocks, nor were they observed to encounter such issues. However, this meant that other issues became more salient. For example, two of the five participants who left feedback highlighted the inability to reorder survey questions. One participant explained:

"When designing surveys, movisensXS allows you to change the order of questions after they have been written - useful if you make mistakes and dont want to type everything out again"

In this iteration of Jeeves, survey questions could not be re-ordered, such that any mistakes involved deleting and re-adding questions - a process with high *"viscosity"*, in relation to the Cognitive Dimensions. This issue, while clearly critical for researchers designing studies with large surveys, was not mentioned by any participants in the first study, who instead focused exclusively on viscosity issues with manipulating the blocks themselves.

Resolution - The resolution of this issue simply involved adding the ability to reorder questions in Jeeves. However, this also meant that constraints had to be managed; specifically, questions conditional on answers to previous questions had to be in consecutive order, as shown in Figure 6.14.

6.3.5.2 Issue 2 - Cognitive Load

From observing participants completing the tasks, it was noted that they spent time shuffling triggers around the canvas upon adding or removing other triggers. Moreover, two participants suggested that the inability to organise triggers was detrimental to their experience. One participant commented: *"I like how the blocks in movisensXS snap into place - makes it easier to have a neat layout"*, with another desiring the *"ability to add triggers together to tidy look of visual UI up"*.

Further, one participant made a comment indicative of the *"hard mental operations"* dimension:

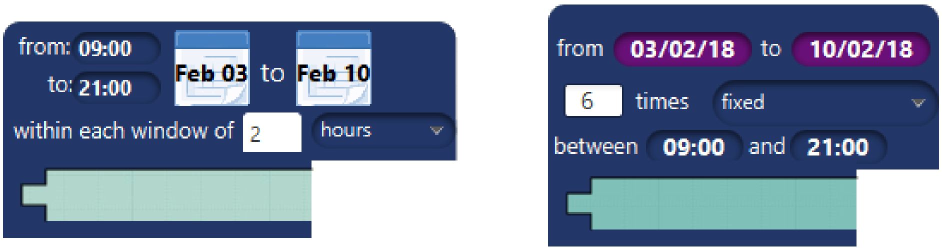


Figure 6.15: Old (left) and new (right) repeated time triggers: window time replaced by prompt frequency

"Jeeves only lets you select windows of '__' minutes, but movisensXS allows you to select a 'number of prompts' per time selected, which may be more useful". This is likely due to the phrasing of the studies replicated in Task 2. For example, the study described by Wonderlich et al. specified *"six semi-random signals throughout the waking hours of the day"*. In forcing participants to calculate the length of window that would equate to six prompts, this induced cognitive load.

Resolution - On reviewing further previous studies in the literature, it was decided that specifying the number of prompts was more useful than the window length in which they should occur, thus this was updated as shown in Figure 6.15. The ability to sort or combine triggers was not implemented, and instead constitutes future work.

6.3.5.3 Positive feedback - Jeeves potential

One positive aspect of Jeeves suggested by participant comments was its potential to act as a tool for designing self-monitoring applications. This was strongly represented by one participant's comment:

"Personally, if there could be a way for the user to view his/her own information regarding emotion regulation, I think that would be very useful. I think it's really useful to keep track of personal behaviors that trigger depressive states, and Jeeves provides an easy, hassle free way of keeping observational records"

Given that participants were only expected to comment on the ease-of-use of Jeeves, it was promising to receive feedback on its perceived usefulness. Another participant similarly mentioned a potential feature to *"link users to webpages or phone emergency services from the interface"*. Recognising the potential fragility of patients involved in clinical psychology, this highlighted the need for Jeeves to act as a tool for providing participants or patients with tailored interventions.

6.3.6 Discussion

This workshop study confirmed that the major usability issues identified in Study #1 had been addressed, and also gave a preliminary insight into the practical application of Jeeves in health psychology research and practice. Perhaps the most important result of this workshop was participants' perceived experience of the comparative evaluation, strongly suggesting that this type of comparison is inappropriate for evaluating Jeeves.

Participants did not have equal exposure to *MovisensXS* and Jeeves, and only had a brief explanation of *MovisensXS* in the form of its built-in interactive tutorial. Consequently, much of their feedback claimed that Jeeves was easier to use and more understandable. For example, participants describing *MovisensXS* wrote: *"I found it incredibly complicated. I wasn't sure how to use it at all and got very frustrated as nothing seemed to work for me"* and *"more difficult, less intuitive, little explanation as to what the actions did"*. Yet, customer testimonials of *MovisensXS* suggest that researchers do in fact find it intuitive and easy to use[1]. As such, it appears that either tool can be seen as "easy to use" if adequate support is provided.

Although an empirical usability comparison could give suggestions for minor usability improvements, the most *useful* EUD tool from an end-user's perspective is not that which minimises time taken and error rate in accomplishing specific tasks, but which allows the end-user to implement desired functionality that would otherwise require programmer involvement. Furthermore, designing study tasks for a fair comparison would require that the chosen tasks could be completed with either tool. Given that the shared feature set of *MovisensXS* and Jeeves is smaller than the feature set of Jeeves on its own, as shown in Table 6.3, it is impossible to comparatively evaluate the usability of these novel features. Given the necessary introductory time required for each tool, a much richer evaluation of Jeeves could be achieved in the time taken to have participants interact briefly with both Jeeves and *MovisensXS*.

However, a comparative evaluation is still an important area of future work for assessing the real-world value of Jeeves. The usability of shared features would constitute a smaller part of a qualitative comparison involving researchers with an active interest in running an ESM study, who would instead have greater insight into the *usefulness* of either tool. Such a comparison would have researchers developing and deploying a study with both tools, through which it could be discerned whether the blocks-based programming paradigm was the most *usable*, but also whether the unique features of Jeeves are more relevant to the development of *useful* ESM studies. This is further discussed in the Future Work of this thesis (Section 9.4).

[1]https://xs.movisens.com/customers Accessed 17/03/18

Table 6.3: Novel and mutual features of MovisensXS and Jeeves

MovisensXS features	Mutually testable features	Jeeves features
Multimedia input questions	General survey creation	Context-sensitive triggers
Reaction time questions	Compliance monitoring	Survey triggers
Individual weekday triggers	Time-based triggers	Conditional expressions
Minimum prompt separation	Button-based triggers	Participant attributes
etc...		etc...

6.3.6.1 Limitations

This study had some further limitations. For example, in Task 2, it emerged that designing a study specification from a relatively basic description proved difficult for participants. However, it is not clear whether this was due to participants' inexperience with ESM terminology, an issue with the clarity of Jeeves, or simply inadequate description in the research publication. For example, some participants expressed confusion when study descriptions did not provide specific questions and they were asked to conceive their own. To isolate any issues with Jeeves itself, this task would need to be conducted with participants familiar with ESM, provided with specific survey questions to be asked.

Participants' progress through Task 2 was therefore slower than anticipated. Although all were able to complete the guided implementation in Task 1 with little trouble, study replication slowed participants, such that the workshop's structure was changed to allow participants to move on to Task 3 once they had designed Jeeves specifications for the first two published studies only. The difficulty of this task was reflected in one participant's feedback, who said: *"the exercise which was related to 5 different research papers, was boring and it took my interest in workshop"*.

Partly due to the difficulty of Task 2, which limited time towards the end of the study, participants also appeared to struggle with the free-form design task. The constraints of the workshop meant that participants generally had very little time to conceive a design, but it is likely that this was also due to their lack of domain knowledge.

Given this lack of domain knowledge and small number of available participants, it may instead have been more appropriate to conduct a cognitive walkthrough with usability experts [191]. With this method, the expert performs the actions required to complete a task representative of a typical user. At each stage, the expert assesses whether a user would be able to identify the next correct action. While the cognitive walkthrough was considered, it was decided that obtaining issues directly from non-expert users on Jeeves in its entirety would be more valuable than focusing on specific tasks.

6.4 Study #3: Think-aloud and Feature Evaluations

A final usability evaluation was conducted in August 2017, in order to evaluate further implemented features. Additionally, this study sought to address a limitation common to both previous studies - the inability to explicate participants' understanding, and reasons why they encountered particular issues. Although brief interviews were conducted in the first study, feedback relied on participants' recall of their progress through the tasks, which prevented momentary thoughts relating to their issues. Thus, the focus of this evaluation was on the effectiveness and satisfaction dimensions of Jeeves, applying qualitative methods to elicit participant feedback and to clarify specific issues.

Moreover, the previous two studies employed tasks that gave instructions or suggestions for which blocks to use. Having resolved the problems of direct manipulation that participants faced in the first two studies, this study was intended to focus on problems of *cognition* that participants encountered in completing more complex, less guided tasks. The usability of an EUD environment in particular relies upon an accessible representation of the problem domain in addition to standard usability heuristics.

6.4.1 Participants

Prior to running the study as described in this section, a pilot study was conducted with two students in the School of Computer Science, to ensure that the think aloud protocol was manageable for participants while completing study tasks, and to check that the semi-structured interview questions were sufficient to elicit useful feedback. In the main study, participants were 10 students at the University of St Andrews, recruited via advertisement in weekly student memos, and through circulating specific emails to students in the schools of psychology and medicine. The advertisements requested that participants should have an active interest in data collection apps, either for participants in their area of research, or for themselves with respect to self-monitoring. In total, six participants studied psychology, two studied medicine, and two studied international relations, with a mix of undergraduate and postgraduate students. Three psychology postgraduate students reported experience with MATLAB. One of these students had employed paper-based ESM in a previous study as part of his PhD research.

6.4.2 Procedure

Prior to running this study, a 10 minute tutorial video was prepared, which had a number of advantages over previous introductions to Jeeves. The tutorial was designed to guide participants through the necessary information they would need to complete the study tasks, by demonstrating an example app specification being built. This video demonstration was considered to be clearer than the paper instruction booklet, and the GIF animations used in the previous two studies. In addition, rather than having an instructive walkthrough task, as also conducted in the previous studies, this allowed more time for participants to complete more complex, cognitively demanding tasks. Finally, the tutorial video served as a useful reference for when participants were unsure how to proceed.

Unlike the previous two evaluations, participants were instructed to think aloud as they completed their tasks, in order to understand *why* specific issues were encountered, but also to explicate participants' mental models of triggers, conditions and attributes. Although a think-aloud protocol can distract participants from fully concentrating on tasks, and thereby increase time taken, the purpose of this study was not to derive a quantitative measure of efficiency, as in Study #1. As such, time taken and number of errors made were not calculated in this study (although the *type* of errors participants made were still of interest, as discussed in the results section).

6.4.3 Tasks

In the period between this study and Study #2, a large proportion of features in the current iteration of Jeeves were developed. The tasks in this study were thus designed in order to assess these features that had not been evaluated, namely:

1. **Context-sensitive functionality**: New triggers, actions and expressions related to sensor data were added. While participants in previous studies interacted with such blocks in walkthrough exercises, they had not been asked to do so without guidance.
2. **Event and state triggers**: Again, previous study participants had been shown how if-conditions could fit into triggers to distinguish event and state triggers. However, it was unclear whether participants could distinguish the different use cases with no explicit instruction.
3. **Creating and setting attributes**: This study was the first in which participants were asked to create and apply their own attributes with no explicit guidance.
4. **Monitoring compliance**: The 'User Data Pane' for monitoring compliance and downloading data had not previously been introduced in previous studies, hence the comprehension of this component of Jeeves was evaluated for the first time.

Nielsen's guidelines on designing study tasks were followed closely, by ensuring that participants were not primed with the terminology of Jeeves [192]. For example, the concepts of triggers, actions and expressions were not explicitly referred to in task descriptions. The full list of tasks, and the features they were designed to test, are shown in Table 6.4. Note that the hypothetical app end-users are referred to in study tasks as *patients* rather than *participants*, in order to distinguish them from the participants in this study. Further detail of these tasks is given as follows:

6.4.4 Data collection and analysis

In this study, no comparisons were made between levels of participant experience (as in Study #1) or between ESM creation tools (as in the previous study). Instead, each participant was exposed to the same introductory video and asked to complete the tasks in the same order.

As in Study #1, these tasks were completed with one participant at a time, in a room where the researcher was present. All participants used the same iMac machine, which ran screen and audio capture software. Participants' interactions with Jeeves, their think-aloud comments during these interactions, and post-study semi-structured interviews, were all recorded for later analysis.

Efforts were made to reduce response bias, by introducing Jeeves as a tool in its early stage of development, for which both positive and negative feedback would be useful for the researcher. Also, interview questions were phrased to prompt neither positive nor negative responses. Further, procedural bias may have occurred due to the think-aloud protocol: this can distract participants from completing the tasks effectively, and a careful level of researcher prompting is required to ensure that participants continue to talk without becoming uncomfortable - a form of experimenter bias. Further, response bias can occur where participants take time to formulate carefully worded statements, rather than articulating all their thoughts as they occur.

A limitation of the analysis described in this section is that only one researcher performed the coding of participant issues, and this process was not analysed by an independent coder. Armstrong et al. elaborate on the use of inter-rater reliability in qualitative research [193], where they showed that although there is inherent subjectivity in qualitative coding, a general consensus of major themes is agreed upon by a set of experienced researchers. Thus, given the relative inexperience of the researcher in analysing qualitative data, this is regarded as a further limitation of the study analysis.

6.4.4.1 Task 1 - Patient Attribute Usage

This task assessed whether the sequence of creating attributes, setting them via survey questions, and then using these attributes in the blocks specification was comprehensible by participants. Following this, task also determined the means by which participants would identify the correct blocks to utilise implicitly sensed context as both an action *("sample nearby Bluetooth devices")* and a trigger (*"when the patient returns home"*).

Table 6.4: Tasks presented to participants, and the features they were designed to test

Task	Description	Tested Features
1.1	At the very start of the study, get the patient's waking and sleeping times.	Attribute Use Implicit Context Survey Creation
1.2	For two weeks, three times a day while the patient is awake, remind the patient to "Take time to practise your mindfulness exercises"	
1.3	A patient's home environment could be an interesting context to sample stress. Tomorrow morning, ask the patient where they live.	
1.4	It would be useful to know who else lives at home with the patient. Sample nearby Bluetooth devices when the patient returns home.	
2.1	Update your study so that the patient can log an interesting location.	Log Design Implicit Context Survey Creation
2.2	After they log their location, the patient should be asked how stressed they feel, from 1 (not very stressed) to 5 (extremely stressed).	
3.1	Check whether patients have been completing the surveys regularly.	User Data Pane Blocks Readability
3.2	If any patient is having a problem with compliance, ask them if they are having issues with the study.	
3.3	Check whether surveys are being sent too often or not enough.	
3.4	Update the study to ensure that surveys are sent at appropriate times.	
4.1	Jeeves should be able to remember that the patient is in Do Not Disturb Mode when a 'Do Not Disturb' button is pressed.	Log Design Attribute Use Conditionals
4.2	Patients should not receive reminders when they're in Do Not Disturb mode.	

6.4.4.2 Task 2 - Survey Button Creation

Participants were asked to design a specification that would allow a patient to "*log an interesting location*" and subsequently send a survey asking their stress rating. This task assessed whether participants would choose to automatically capture the patient's current location using the *Sense Data* action, or whether they were primed from the previous task to obtain this information through a survey. It was also interesting to observe whether participants would also assume that storing answers in attributes is necessary (it is not necessary, as all survey answers are uploaded to the database, regardless of whether these answers are also attribute values).

6.4.4.3 Task 3 - Patient Compliance Reaction

Participants were asked to load a study configuration that had been populated with dummy patient data. In this study, one patient had failed to complete any surveys, and all patients had missed multiple surveys because the configuration was faulty, designed to send surveys 50 times a day. In addition, an 'end of day survey' was not sent at all because its trigger did not have a time. These two faulty triggers are shown in Figure 6.16. This task assessed the readability of the blocks-based notation, as well as that of the User Data pane, and whether participants could act upon the data presented.

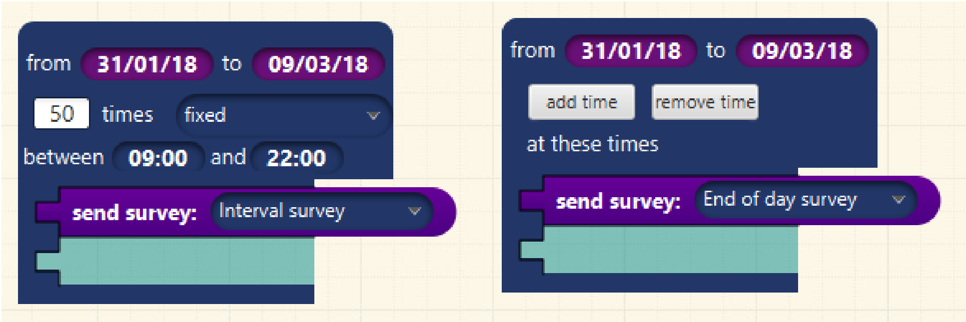


Figure 6.16: Two of the 'faulty' triggers in the Task 3 specification

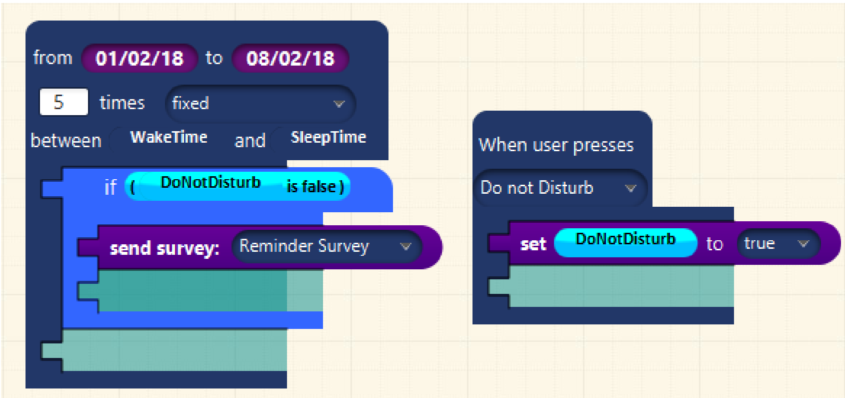


Figure 6.17: An example 'Do Not Disturb' specification

6.4.4.4 Task 4 - Do Not Disturb Button Creation

Finally, participants were asked to implement a 'Do Not Disturb' button that would stop patients receiving prompts, an example solution of which is shown in Figure 6.17. It was intended that participants create a 'Do Not Disturb' button and attribute. Pressing the button would set the attribute to be true. The reminder survey in the repeated time trigger then had to be nested within an if-condition based on the attribute's value. This was a comparatively difficult task that assessed the complexity of automatically setting attributes, and hybrid event-state triggering. While a similar application was demonstrated in the tutorial video, participants were required to recognise the similarity and adapt it to the task's requirements, as would be expected of researchers in a real-world scenario.

6.4.5 Usability issues analysis

Through the think-aloud participant feedback and post-study interviews, this study obtained the richest source of qualitative data, in order to address whether participants understood potentially difficult concepts in Jeeves, such as attributes, conditional expressions, and context-awareness.

The study provided insights into participants' mental models of how they believed blocks should be combined, and their ability to recognise and recall specific usage patterns. Participants' think-aloud monologues and post-study interviews were transcribed, and a thematic analysis of their issues with Jeeves was conducted. Further, all issues that participants encountered were coded, based largely on the learning barriers in EUD environments proposed by Ko, Myers and Aung [194]. A brief description of these barriers follows:

Six Learning Barriers in EUD environments

- **Design barriers** are those where a user cannot formulate a programming problem into an appropriate design to solve that problem.
- **Selection barriers** occur after design barriers, where a user has decided that they know how a program should be designed, but do not know which constructs to use.
- **Use barriers** are those where a user has selected a construct that they believe should be used, but is unsure how exactly to use it.
- **Coordination barriers** are encountered when a user knows the set of constructs to use, but is unsure how to make them work together.
- **Understanding barriers** are those where a program has unexpected behaviour, and the user is unsure why this behaviour has occurred.
- **Information barriers** refer to instances where a user is unable to check their hypothesis of why their program functioned incorrectly.

This was a particularly insightful analysis that was not possible in the previous study, due to lack of screen recording, nor the first study, as the tasks were too specific and did not encourage participants to employ problem solving skills. Moreover, the majority of issues encountered in Study #1 were not learning barriers, but direct manipulation barriers. Assessing learnability is important from an acceptance perspective, since an interface that cannot be learnt easily will likely be abandoned in favour of alternative means [146].

According to Bau et al. [138], blocks-based languages can reduce *selection*, *use* and *coordination* barriers. *Selection barriers* are reduced because blocks are visible in their respective palettes, enabling *"recognition over recall"*. *Use barriers* are reduced because textual code constructs are combined into more meaningful "chunks". Finally, *coordination barriers* are reduced because the secondary notation of blocks affords their correct combination.

While these features of Jeeves generally assisted in completion of tasks in this study, such barriers were not alleviated altogether, and indeed participants occasionally encountered fundamental design barriers. The issues that participants encountered in their tasks are described below in

Table 6.5
Participant backgrounds and their completion of study tasks

PID	Background	Programming	Task 1	Task 2	Task 3	Task 4
P1	Psychology	R, MATLAB	🟡	🟢	🟡	🟢
P2	Psychology	MATLAB, Python	🟢	🟢	🟢	🟢
P3	Medicine		🔴	🟢	🟡	🔴
P4	Psychology		🟡	🟢	🟡	🟢
P5	Medicine		🟡	🟢	🟢	🔴
P6	IR		🟡	🟢	🟢	🟢
P7	Psychology		🟡	🟢	🟢	🟢
P8	Psychology	MATLAB	🔴	🟡	🟡	🔴
P9	Psychology	HTML/CSS	🟢	🟢	🟡	🟡
P10	IR		🔴	🟡	🟡	🔴

🟢=Completed 🟡=Completed but with errors 🔴=Incomplete, insurmountable barrier

chronological task order, with potential resolutions of these issues. Relevant backgrounds of participants, and their completion of study tasks, are shown in Table 6.5.

6.4.5.1 Task 1

This task assessed a broad range of functionality, including attribute usage, survey creation, and use of context-related blocks. As such, many issues were encountered that led to major *Design Barriers* for two participants.

Issue 1 - Hidden dependencies

Creating and assigning attributes in Tasks 1.1 and 1.3 caused a surprising number of *Coordination Barriers* to participants. The sequence of creating an attribute, creating a survey, assigning the attribute to the survey, and then sending the survey, caused issues when participants would try to circumvent this process. Participants frequently dragged attributes into the "Send Survey" action, or directly into the triggers themselves. This suggests that the approach has a high viscosity, in terms of the physical effort required in making a cognitively simple update. The approach also gave rise to *Use Barriers*, where participants were oblivious to the fact that they had missed one of these steps. This instance of *"hidden dependencies"* suggests that a more streamlined means of implementing this process might be necessary, or the dependencies communicated to the user.

Resolution - As described in Section 5.3, the consistency and visibility of Jeeves relies on introducing hidden dependencies, such as that which caused this issue. However, one participant

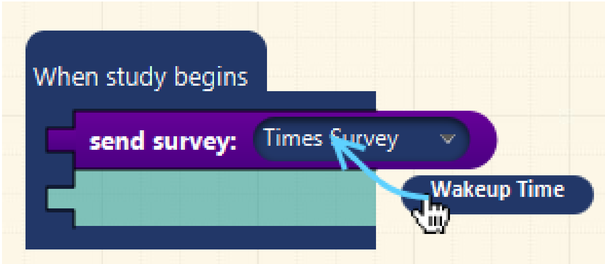


Figure 6.18
Participants wanted to save answers as attributes via dragging and dropping, as shown

suggested that this also introduced an inconsistency, and that the attribute should somehow be linked to a survey question on the blocks canvas, as shown in Figure 6.18:

"Like in this window, you [save answer as an attribute] in this window, in the survey design. But in the end you are using it [on the blocks canvas] so, I mean it transfers but for me it was confusing that you did it here" (P8)

Issue 2 - Too much abstraction

Issues were also caused by unsuitable levels of abstraction. Five participants in Task 1.3 expressed a struggle to find a one-time trigger. While this function could be implemented by adapting a "Set Times Trigger" block, this was unclear to participants, resulting in a *Selection Barrier*. Over-abstraction also resulted in participants experiencing a *Use Barrier* with the "Capture Data" action, desiring further customisation, as expressed by one participant: "*Is that it?...Can I tell it specifically what type of data from Bluetooth I want?"* (P2). However, most participants did not encounter these issues, suggesting that different levels of abstraction are appropriate for particular participants.

P6, who completed tasks quickly and correctly despite his lack of scientific background, further suggested the possibility of adding his own abstractions for complexity management:

"So I have my press commands and maybe group them, or maybe a way to group them on my own, my own volition, maybe you could pull them onto each other and make them into a group" (P6)

Resolution - As suggested by P6, introducing a means for researchers to create their own abstractions could solve this issue, ensuring that only a level of abstraction appropriate to the individual would be available. Alternatively, support could be added to allow experienced researchers to deconstruct high-level abstractions into their base functions.

Issue 3 - Searching by name, not type

Task 1.4 appeared to cause issues for all participants. Bau et al. suggest that organising blocks

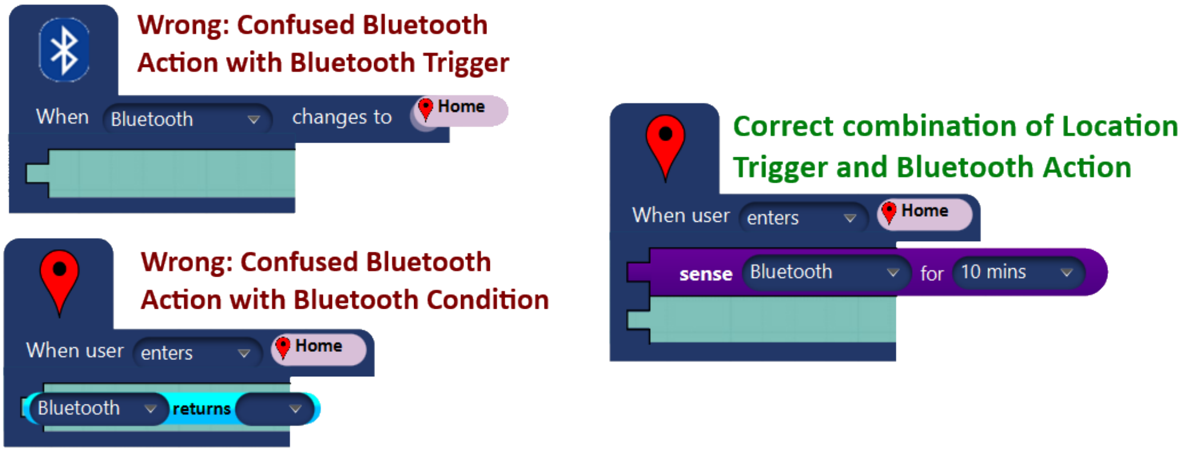


Figure 6.19
Language ambiguity caused trigger, action and condition confusion

by type into palettes should support *recognition over recall* [138]. However, participants' inexperience with the concepts of triggers, actions and expressions meant that in many instances, they would not consider the correct palette in which to find certain functionality. Instead, they would click at random through all palettes, hunting for a block by name, regardless of its type. Figure 6.19 shows two examples of participants failing to find the appropriate "Sense Data" action - instead misusing the sensor trigger, or misusing a sensor condition in place of an action. When participants eventually found and used the correct action, they expressed doubt as to correctness. Moody argues for using text to complement graphics:

"When information is presented both verbally and visually, representations of that information are encoded in separate systems in working memory and referential connections between the two are strengthened" [162, p. 771]

Green and Petre further suggest that text can support graphics as a form of secondary notation [161]. However, in Jeeves, where graphics are used to communicate a block's grammatical properties, and text is used for showing semantics, this had the opposite effect. Participants often did not consider the 'visual syntax' required, and instead tried to use the first block that had a similar semantic meaning.

Resolution - Researchers may not initially think in terms of triggers and actions. The feedback received from participants suggest that graphics should not take precedence over text in distinguishing meaning. Grouping blocks by their semantic properties (such as sensor use, times, or locations) rather than their syntactic type, could resolve this issue, although investigating this constitutes future work.

6.4.5.2 Task 2

All participants completed this task with no *Design Barriers*, in that all provided a reasonable solution. Creating a button and specifying its behaviour through a trigger, in order to implement event-contingent sampling, appeared to be an intuitive process. Notably, *Use Barriers* were encountered that, while not critical to the correctness of the specification, nonetheless showed that participants had incorrect assumptions of how Jeeves would work in practice.

Issue 4 - Unnecessary Questions

As an example of incorrect assumptions, five of the ten participants opted to use a survey to acquire patients' current location, rather than capturing this automatically. This was not related to an inability to find the "Sense Data" action in the first task, and instead was due to ambiguity of the action's function. Participants were unsure whether patients' location data would be synchronised with their survey responses, and expressed doubts as to the difference between capturing location data automatically, and storing a patient's chosen location in an attribute:

"When user presses log they can add a location. Capture data from...I'll put a new attribute, 'location'. Is that right? Capture data from location, right? What's the difference between 'this' location and the location attributes?" (P4)

Issue 5 - Unnecessary Attributes

Additionally, it was observed that four participants, with no indication of uncertainty, created unused attributes to save patients' answers to survey questions. Their erroneous assumption was that answers had to be saved as attributes, in order to be accessible to the researcher. As described by P1: "*It's quite difficult to distinguish this patient attribute from the data we're trying to measure*". From this observation, it appeared that the usage pattern of assigning an attribute to a survey question in the previous task primed participants to assume that this was necessary for all survey questions.

"Do I need a new attribute? If it's just going to be the answer to a question, if I'm not going to use that answer then it doesn't need to be an attribute. Or does it? Okay, uh, I'll create an attribute anyway" (P2)

Resolution– The resolution of both the above issues appears to be a matter of clarifying the process of data storage. Because participants were unaware of how sensor data was stored in the database, and whether saving data values into attributes affected this storage, they frequently created unnecessary attributes and questions. A possible solution would be to show a live representation of the database's structure as a study is updated, or simply to describe data storage in a tutorial video.

6.4.5.3 Task 3

Task 3 also caused few identifiable issues with participants. The User Data pane was designed as a GUI separate from, but simultaneously visible with, the blocks-based specification, with the intention that issues with compliance identified in the User Data pane could be related to a faulty study specification.

No issues were caused with participants viewing the data of patients and discerning those that were particularly non-compliant. However, in updating the study to address this non-compliance, different approaches were taken. All participants discovered that one survey was being sent 50 times a day, but only four found that the end-of-day survey was not being sent. This presents an additional issue, described as follows.

Issue 6 - Gulf of Evaluation

There was a prevalent *Information Barrier* for participants in completing this task. Simultaneous visibility of the User Data pane and the blocks specification was insufficient for participants to make explicit connections between the two. This suggests the presence of a *"Gulf of Evaluation"*, defined by Donald Norman as follows:

Slips and Mistakes

"The Gulf of Evaluation reflects the amount of effort that the person must make to interpret the physical state of the device and to determine how well the expectations and intentions have been met" [195, p. 39]

As Jeeves does not currently provide feedback on the function of created specifications, participants expressed doubt as to the correctness of their solutions to tasks, or were unaware of subtle mistakes that could cause faults in their developed apps.

Interestingly, regarding the 'faulty' trigger designed to send participants surveys 50 times a day, two of the psychology postgraduate participants opted to not reduce this number of daily surveys, One explained: *"I don't know how many data points the researcher intends to collect, so I'm gonna leave 50 there"* (P2). Regardless of the number of data points desired, previous research has shown that there are upper limits on the number of surveys that study participants can be reasonably expected to comply with. For example, studies signalling participants eight or more times a day over the course of two weeks have compliance ranging from 50-70% [8].

Resolution - An obvious resolution to this issue would be to incorporate a simulation of live execution into Jeeves, similar to the "live programming" feature employed in *Scratch*. However,

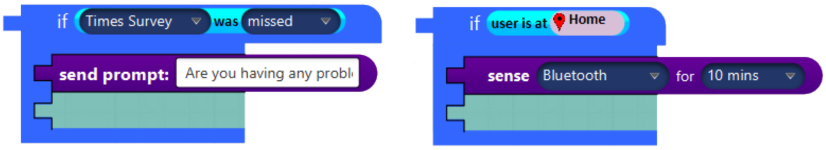


Figure 6.20
If-conditions were misinterpreted as event-based triggers by P2 (left) and P10 (right)

as an app's behaviour is contingent on context, it is not clear how such a simulation could be applied by researchers, and may be a complex EUD task in itself if researchers have to specify variations in context and observe how these affect their app's function. An alternative solution could be to include some form of intelligent recommender system to advise against particularly intrusive study specifications. Haines et al. suggest the use of recommender systems for supporting end-user development [196].

6.4.5.4 Task 4

This task cased the majority of issues for participants in attempting to combine event and state triggers. Previous research has shown that even in a blocks-based visual programming paradigm, different trigger types can still be confused if they are not visually distinct [168]. Unambiguous context-aware behaviours should be composed of one discrete *event* trigger, with an optional number of *state* conditions [159, 197]. As shown in Figure 6.19, Jeeves, as with other blocks-based languages, supports these constraints through the visual affordance of correct connections. Event-based triggers cannot be combined, for example, as their shapes do not nest.

Issue 7 - Events and States

As expected from prior work by Huang and Cacmak, despite the visual distinction, participants had trouble separating triggers involving discrete, instantaneous events, and continuous, ongoing states [197], such that combining the repeated time event with the 'Do not Disturb' state introduced *Use barriers* and *Coordination barriers* for most participants. For example, P2 initially assumed that an if-condition detached from a trigger would enable a change in the value of an attribute to be detected like a discrete event. P10 also made this assumption, with their examples shown in Figure 6.20.

"I'm going to assume that this (if-condition) will be continuously running and that it doesn't need to be attached to a trigger object" (P2)

"Anyways let's do this one. IF location changes, capture data from Bluetooth!" (P10)

Although the if-condition block's external connector affords nesting within a trigger, this was

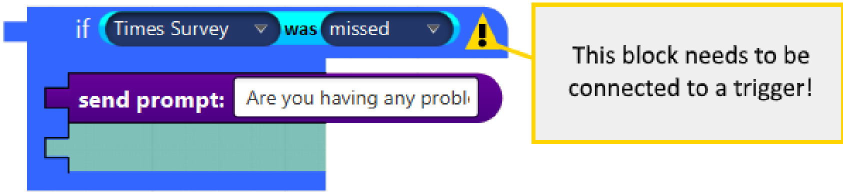


Figure 6.21
A possible resolution using additional secondary notation

insufficient secondary notation to ensure that participants would do so. Thus, similar to erroneous attempts to join incompatible blocks, participants' mental models influenced their decisions regardless of secondary notation.

Resolution - Participants were unsure how their mental models could have been corrected. One possibility would be to incorporate additional secondary notation on blocks that are disconnected. *App Inventor*, for example, displays a warning triangle on disconnected blocks, which reveals a pop-up with an explanation. Huang and Cacmak also suggest making use of prompts for incorrect specifications [197]. An example of this is shown in Figure 6.21.

Issue 8 - False Coordination

Coordination Barriers were generally overcome within a few seconds, when participants discovered that blocks did not connect together. However, inaccuracy in dropping blocks caused insurmountable 'false' Coordination Barriers. Figure 6.22 (right) illustrates how a 'near miss' had a negative impact on participants' mental models of how blocks could be combined. P5 who made this error, stated she was unable to understand how to use the block that had caused this confusion. When explained to her, she described the issue:

"I just saw it again in the video, you need to drop the thing inside but like it sometimes doesn't go in, so you feel like it just hovers so it doesn't always go inside the space" (P5)

Participants learn that a block snapping into place indicates correctness, so when this fails to happen due to an inaccurate dropping action, correct choices are discounted and mental models break down. Norman distinguishes *slips* and *mistakes* as follows:

Slips and Mistakes
"Errors have two major forms. Slips occur when the goal is correct, but the required actions are not done properly: the execution is flawed. Mistakes occur when the goal or plan is wrong." [195, p. 170].

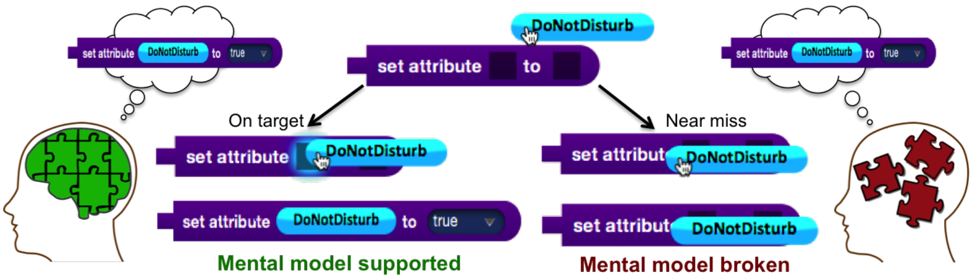


Figure 6.22
Left: A correct snap supports a user's mental model. Right: A 'near miss' damages the mental model

As such, direct manipulation of blocks can cause slips, which could further lead to more critical mistakes if these slips are not addressed immediately.

Resolution - These slips, while not learning barriers of Ko et al. in themselves, were unresolved direct manipulation issues that thus caused mental models to break down, and induce such learning barriers. One possible resolution is expanding the drop target in response to the cursor's proximity, as proposed by Cockburn and Firth [198]. This simple exploitation of Fitts' Law was shown to improve target acquisition and was popular with study participants. An illustration of how such a feature would look in Jeeves is shown in Figure 6.23.

Issue 9 - Lack of abstraction

While too much abstraction caused barriers in Task 1, barriers in this task were conversely caused by a *lack* of abstraction. For example, five participants indicated that they were searching for a 'Do not Disturb' action, rather than having to implement this functionality themselves. This was expressed by P5 in her interview:

"I realise that a lot of the things I wanted to do could be composed of the triggers and actions and conditions themselves but I guess for me, they'd be kind of like, simpler options? Or simple components?" (P5)

One example of this can be seen in the TouchDevelop application [199], which aims to provide the "gentle slope" of complexity aimed for in EUD tools [16], by providing three different modes of input, beginning with a high-level blocks specification and ending with users engaging in lower-level scripting.

Resolution - As suggested in resolving Issue 2 (too much abstraction), further research should establish with researchers and clinicians what useful fundamental blocks should be, so that commonly employed features do not have to be manually constructed from more basic blocks every time they are required. Once useful abstractions have been established through iterative

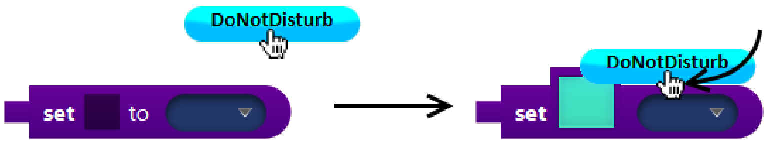


Figure 6.23
Improving target acquisition by expanding drop target

design, these could then be broken down by skilled researchers to form new abstractions. Moody's Principle of Semiotic Clarity states that *"there should be a 1:1 correspondence between semantic constructs and graphical symbols"*. However, given that different researchers may think of the same domain in conceptually different ways, this principle has to be flexible in its implementation.

Issue 10 - Ambiguous Actions

Finally, the apparent ambiguity of the previously implemented "Wait" action was a particular source of error, causing *Use Barriers* with many participants. This action was intended to pause execution of subsequent actions in a particular trigger. However, in trying to complete the task of implementing a 'Do Not Disturb' button, participants assumed that this action would pause notifications across the entire application (Figure 6.24, left), showing a misunderstanding of trigger concurrency. Additionally, three participants attempted to use their 'Do Not Disturb' attribute with this action, in an attempt to wait until the attribute was false (Figure 6.24, right). These participants also assumed that state changes were events themselves.

Resolution– Participants suggested that the word "wait" in this context was ambiguous and needed further explanation. When asked what format this explanation could take, participants suggested providing examples, which could be built upon:

"Say I saved this and I wanted to send a similar one out, this could pop up again. Then I could change it a little bit and re-save it as something else" (P3)

Providing researchers with a library of examples would help resolve the *Design Barriers* that participants encountered, as suggested as a supporting EUD approach in Section 4.1. Moreover, an end-user community where "power users" construct specifications for non-programmer users to build upon is a particularly beneficial product of the meta-design approach.

Participants were enthusiastic about the value of having a tutorial video. However, some expressed their preference for learning by practical application, and desired support to explore undiscovered functionality, through more targeted tutorials:

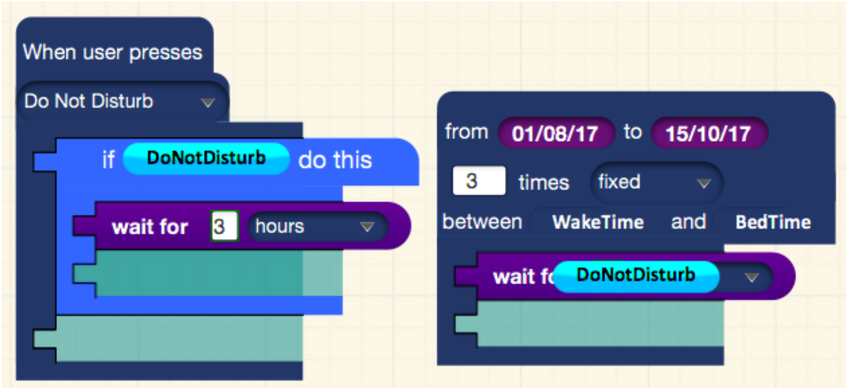


Figure 6.24
Two erroneous attempts at 'Do Not Disturb' in Task 4

"Maybe like a video at each [block], like Begin Trigger, and it shows how you could use Begin Trigger. like it can be very short like 30 seconds" (P7)

6.4.6 Positive features analysis

Although the number of issues identified is comparatively higher than those of the two previous studies, it is proportional to the depth of this study in its efforts to elicit participant feedback. As such, participants provided a wealth of positive feedback related to usability, as well as potential real-world application. In terms of usability, the benefits of the visual metaphor, and the approachable learning curve, were prevalent in participants' comments.

6.4.6.1 Positive Features - Visual Metaphor

Participants highlighted the benefits of the jigsaw-puzzle metaphor in reducing the effort required to complete the study tasks. The secondary notation of the blocks in terms of shape, colour and size meant that some participants were able to make the connection between text and graphics quickly:

"I guess if something was not working you'd be able to figure out what you'd put in that wasn't right, 'cause its not all the same colour and your brain's already associating this with the main thing and that with something else" (P3)

As well as the benefits of visual metaphors in themselves, one participant, who had previously used a diagrammatic visual language, expressed the benefits of the blocks paradigm over this alternative flowchart representation:

"A flowchart makes you feel like you're trying to think like a computer, like you have to know

what you're doing from the start, whereas this is good because you can go like 'oh this is wrong I'll add this'" (P6)

This feedback supports the idea of *"tinkerability"* proposed by the authors of *Scratch* [200], or programming by *"bricolage"* [164], whereby uses observe the elements at their disposal and rearrange them to fit their ideas, rather than employ a structured software development approach. This feedback is in line with the results of Desolda et al. in their study of composition paradigms for trigger-action rules, showing that not imposing an order of creation improves users' performance and satisfaction [125].

6.4.6.2 Positive Features - Learnability

As in Study #1, the majority of participants stated that Jeeves was learnable and that familiarity would require minimal time. This was a promising result, and as shown in Table 6.5, there were few insurmountable barriers in task completion, particularly given that the relative complexity of the tasks was high compared to the average experience sampling study (as discussed in Section 2.2). Desolda et al. also discovered that their non-expert study participants were capable of constructing complex trigger-action rules, observing that *"end users can master more complex rules if adequate interaction paradigms are provided"* [125].

However, the rate of *how* quickly an end-user could learn to apply Jeeves is less significant than the benefits of mastery. P5, a medicine postgraduate student, explained that 'walk-up-and-use' functionality is not a requirement for particularly useful software:

"Like SPSS, you have to use YouTube videos or you have to go on a course. It's not unusual for researchers to be used to having to do tutorials, classes, sessions..." (P5)

P6, who had previously used a diagrammatic visual language but was unfamiliar with other visual languages, recognised the potential of the blocks-based paradigm for its learnable qualities:

"You might have a way to spin this out into educational software, 'cause this is making you do logic without really realising you're doing it" (P6)

6.4.6.3 Positive Features - Patient Independence

Feedback on the potential real-world utility of Jeeves was elicited from participants with backgrounds in medicine and psychology. The two participants from a medical background were particularly enthusiastic about the potential uses of Jeeves. For example, P5 discussed the importance of shared decision-making in the application of self-monitoring technology:

"...it's important to work WITH participants when you're collecting data so I think being able to

personalise and tailor things means there's less chance of you sending people stuff that they'll find intrusive or annoying or it makes people feel more in control" (P5)

P3 explained that, while it is important that patients are able to self-manage their conditions independently, they may also value the feedback and reassurance that their data is being looked at by a healthcare professional:

"If [patients] needed to record their blood pressure and there was an option to send it back or blood glucose if they're newly diabetic they'd know how to do it and they could send it back, and as a doctor you could go on and check. You'd then be able to realise if something was up" (P3)

6.4.6.4 Positive Features - Lived Experience

One assumption underlying EUD-ESM tools in Section 4.2 is that researchers' goal is longitudinal data collection from participants. However, this does not take into account the potential value of ESM for participants themselves, who have their own model of technology acceptance that determines whether or not they will comply with a study, or engage in self-monitoring. In their design of an app for self-tracking, Karkar et al. suggest that compromising scientific rigour in favour of lived experience would encourage adoption of self-monitoring technology for different purposes than those previously conceived by developers [201]. Supporting this, P9 suggested that her father might be able to design an app for his personal use:

"My dad always forgets to take his medicine, he has diabetes, so I think this would be good to like remind him and then he can take his blood pressure...he always forgets if he's taken it or not and then he could enter 'you need to measure your blood pressure now' and then enter the data"

Furthermore, the usability and general-purpose features of Jeeves encouraged participants to consider its application different domains. One participant, a PhD student in psychology who had previously employed paper-based ESM, was interested in simulating "lived experiences" of disaster situations:

"I expect to collect data from something like this because it's stupid to ask people 'Imagine you were in an earthquake'...I collect a lot of stories, there are a lot of stories that happen on the first day, the second day, the third day. and maybe I can use this story to create an app like this...get participants to receive these stories at particular time" (P4)

Table 6.6
Triangulation of derived design factors with previous design guidelines

Design factor (User Studies)	13 Design Guidelines for EUD [202]	Cognitive Dimensions Framework [161]	Physics of Notation [162]
Employ blocks programming (#1,2,3)	Use objects as language elements Make syntactic errors impossible	Secondary notation Error-proneness	Semantic Transparency Dual Coding
Use text to distinguish graphics(#3)		Secondary notation Diffuseness Closeness of mapping ✗	Visual Expressiveness ✗ Dual Coding ✗
Minimise sequences of actions (#1,2,3)		Hard mental ops Hidden dependencies Viscosity Consistency ✗	
Make syntax correctness easier (#1,3)	Make syntactic errors hard	Error-proneness Visibility Hidden dependencies Secondary notation	
Maximise blocks visible (#1,3)	Multiple views with incremental disclosure	Visibility Diffuseness Closeness of mapping ✗	Graphic Economy Semiotic Clarity ✗ Visual Expressiveness ✗
Support 'tidying up' features (#2,3)		Viscosity Premature commitment	Complexity Mgmt
Support different levels of abstraction (#2,3)	Multiple views with incremental disclosure	Abstraction Visibility ✗	Complexity Mgmt Cognitive Fit Semiotic Clarity ✗
Have modular tutorial videos (#3)	Scaffold typical designs		
Allow reuse of examples (#3)	Scaffold typical designs Build community tools		
Bridge the Gulf of Evaluation (#1,3)	Incremental development Allow Immersion	Progressive evaluation	

6.5 Chapter Summary

Three usability evaluations were conducted with different groups of students in the iterative development of Jeeves. As the complexity of apps that could be created with Jeeves increased, these evaluations served as interim checks that the tool was still comprehensible by those with little or no programming experience. Indeed, the final evaluation showed that participants, with just a 10 minute introduction, were able to complete tasks representative of those required for context-sensitive, individually tailored ESM apps. While usability issues were encountered throughout, these were iteratively corrected in order to move towards *"getting the design right"* [203]. However, the question remains as to whether Jeeves is *"getting the right design"*. Having isolated and removed the user interface problems in the first two studies, learning barriers still persisted, suggesting that there may be a more intuitive representation of an ESM app specification.

In Study #1, the salient usability issues would have overshadowed usefulness in a real-world deployment, leading to frustration and abandonment before adoption. The second study, however, prompted the question of whether immediate usability is necessary. *MovisensXS* is a widely used application in research, with testimonies from researchers as to its ease of use. Yet, participants in Study #2 expressed dislike of *MovisensXS* in comparison to Jeeves. It is not fair to say that Jeeves is objectively *"more intuitive"* or *"easier to use"* as suggested by participants. However, with a longer introduction and more examples provided of using Jeeves, it appeared that usability was simply a case of education and tutorials, such that for two tools providing the same functions, specific interface design factors are insignificant in the presence of adequate support.

In light of this, Study #3 did not attempt to make any empirical comparison with alternative interfaces. Instead, this study focused on the usability of unique features of Jeeves that enabled aspects of study complexity unavailable in alternative EUD-ESM tools. Given that the perceived usefulness of an EUD tool relies on allowing end-users to develop applications previously outside their capabilities, ensuring the usability of a novel feature set is more important than comparing two otherwise equal tools. While perceived usefulness may overshadow ease-of-use in software, as suggested by Greenberg and Buxton [143], a quote by one participant in Study #3 communicates the ongoing need to keep ease-of-use above a certain threshold:

> *"There is a frustration tolerance. You run the risk that people like me would get to this phase and go 'y'know what? I don't know how to do this. I'm just gonna email surveys through Qualtrics because I know what to do'."* (P5)

Thus, usability is still a high priority for meta-designers, to prevent researchers from defaulting to familiar software at the expense of sacrificing potentially useful functionality.

In answering the research question *"How can the development effort be reduced to allow rapid creation of experience sampling apps with no programming?"* it was promising to observe that the majority of implementation decisions made in Chapter 5 were indeed beneficial to end-users in ESM app creation. This summary details the design factors that allow end-users to do so with minimal effort. Some design factors were derived from the final usability evaluation, and thus were not directly implemented and tested in this thesis, but instead provide avenues for future work. These factors are likely to apply to other domain-specific visual programming environments, but verifying such a claim would require application and testing of these factors in different environments. Table 6.6 triangulates these factors derived from this chapter's usability studies with design principles and guidelines from three other sources. In particular, Repenning and Ioannidou's 13 design guidelines for EUD [202], the Cognitive Dimensions of Notation [161] and the Physics of Notations [162] were used. Text with a tick represents the design principle or guideline that supports the corresponding design factor, and text with a cross represents those which the design factor contradicts. This is not to suggest that these principles are wrong. Indeed, their authors acknowledge that certain principles may contradict one another in specific situations. Empty cells in the table are due to these guidelines and frameworks having been developed for different purposes. The Cognitive Dimensions, for example, are for *constructing* visual programs, whereas the Physics of Notations are for *reading* visual programs. Both of these explicitly refer to visual notations, unlike the "13 Design Guidelines" by Repenning and Ioannidou. Because EUD approaches may not incorporate graphics, the derived design factor of *"use text to distinguish graphics"* does not triangulate with any of these 13 guidelines. Further, Repenning and Ioannidou's Design Guideline of *"Scaffold typical designs"* refers to a supporting approach and does not triangulate with factors related to program creation and interpretation.

The 13 Design Guidelines defined by Repenning and Ioannidou do not conflict with one another, unlike the Cognitive Dimensions and Physics of Notations. As such, while Jeeves does not explicitly make use of all 13 guidelines, it does not conflict with any, and indeed future work could investigate the effect of incorporating more of these guidelines. However, the crossed items represent components of the Cognitive Dimensions and the Physics of Notations that conflict.

The factors of *"use text to distinguish graphics"* and *"maximise blocks visible"* both conflict with the *"closeness of mapping"* Cognitive Dimension and the Principle of Visual Expressiveness. In using text to distinguish otherwise identical visual constructs, this reduces the full capabilities of visual variables. Further, when blocks with different functions are only distinguished with text, their appearance does not immediately suggest their meaning, defying the Principle of Visual Expressiveness. In maximising block visibility, this requires having some problem entities represented in other ways, thereby conflicting with *"closeness of mapping"*. Also, because blocks with similar functionality are grouped together in palettes to maximise visibility, this defies the need to maximise the use of visual variables.

A final point of conflict regards the Cognitive Dimension of Visibility that, while supported by two of the derived design factors, also *conflicts* with the factor of *"support different levels of abstraction"*. As abstraction is not currently implemented, future work should investigate whether abstraction can be added in such a way that individual block visibility is not compromised.

- **Employ blocks-based programming:** Arguably this factor is biased, as alternative interfaces were not presented to participants in two of the three studies. However, participants in all studies confirmed that the blocks-based paradigm enabled rapid learning and understanding.
- **Use text to distinguish graphics:** Moody's principle of using text to complement graphics is applicable to readability. However, in using Jeeves to construct a specification, participants instinctively searched for keywords such as 'sensor', or 'survey'. The visual syntax supports the semantic meaning, but these semantics should be distinct and unambiguous.
- **Minimise sequences of actions:** In Study #1, participants strongly disliked the 'click, click, drag' approach for finding and placing blocks. Further, participants were confused by the sequence of sending surveys to capture an attribute. Where possible, it appears that consistency should be sacrificed in favour of supporting end-users' mental models.
- **Make syntactic correctness easier:** Two suggested EUD guidelines, namely: *"make syntactic errors hard"* and *"make syntactic errors impossible"*, could be extended to state that syntactic *correctness* should be *easier*. Although the notation of Jeeves prevents incorrect combinations, participants continuously tried to force them. Extra notation such as highlighting appropriate places for a block could be applied.
- **Maximise visible blocks:** The visibility of blocks was a salient issue in Study #1. Having to check three individual menus to find an action was a major source of frustration for participants. Though this was corrected for Study #2 and #3, participants still claimed they *"forgot about conditions"*, for example. The upper limit of this factor needs to be formally evaluated, but these studies suggest maximising block visibility.
- **Have 'tidying up' features:** Analogous to code clean-up or auto-indent features found in professional IDEs, participants in all three studies expressed a desire for a feature to organise their blocks on the canvas. The unstructured approach to dragging and dropping reduced *premature commitment*. However, participants considered that sorting triggers chronologically, or grouping triggers by their type, would support readability.
- **Support different levels of abstraction:** This factor is closely related to the aforementioned "gentle slope of complexity". Some participants wanted provision of larger blocks to abstract over common functions. However, other participants also suggested the ability to create their

own abstractions, as potential “power users”. Thus, decomposable abstractions would support end-users of all skill levels.

- **Have modular tutorial videos:** Provision of information to participants went through phases, from text and images in a paper booklet, to text and animated GIFs on a website, to a tutorial video. The final approach was unanimously popular, but some participants suggested that shorter tutorial videos, targeted to specific functions, would be easier to navigate.
- **Allow reuse of examples:** Participants valued the tutorial video for learning by example. Participants in both Study #1 and #3 explicitly suggested pre-created examples for supporting learning and efficiency.
- **Bridge the gulf of evaluation:** Participants frequently doubted that they had completed tasks fully, expressing concerns that they had omitted necessary functionality. A rapid evaluation feature is thus of primary importance for future work.

This final guideline derived from Study #3, **Bridge the gulf of evaluation**, is a particularly important finding with respect to the real-world applicability of Jeeves. In both the domains of psychology research and clinical practice, confidence in the reliability of developed apps, both by the individual researcher or clinician, and their organisation, is a prerequisite to adoption (discussed in detail in the following chapter). The ability to instantly test the output of a study specification from a participant’s perspective is an example of Schön’s “reflection in action”, whereby practitioners reflect on their actions in the process of carrying them out [204]. However, of particular significance is that this trial-and-error process could be enacted without direct participant involvement, minimising risk to these participants, as well as recruitment costs for running pilot studies.

It was important to not confuse usability engineering with participatory design [203]. Indeed, usability evaluations should be employed for identifying problems, and not solutions. Asking participants what they would change about an interface, as shown by Tohidi et al., explicitly forces the participant into the role of a designer, for which they do not have the necessary experience. Instead, when issues were discussed in post-study interviews, participants were not asked what they would change about Jeeves, but how they could have been helped in reaching the correct solution.

CHAPTER SEVEN

EFFECTIVENESS ANALYSES

Ideas in this chapter have been published in the following papers:

D. Rough and A. Quigley, "End-user development in social psychology research: factors for adoption," in *Visual Languages and Human-Centric Computing (VL/HCC), 2018 IEEE Symposium on*. IEEE, 2018.

D. Rough and A. Quigley, "Towards end-user development for chronic disease management," in *Designing Technologies to Support Human Problem Solving*. 2018.

Having evaluated the usability of Jeeves with participants of varying technical expertise, a further, and potentially more critical prerequisite for adoption, is its usefulness as observed by potential stakeholders. As discussed in Section 4.1, most research in EUD has consisted of engineering prototype systems and evaluating them in artificial lab settings. However, the organisational and contextual factors of an EUD environment are of primary importance in addressing its real-world practicality. Thus, lab-based usability evaluation is not sufficient for addressing the research problems of this chapter, which are **understanding** the current practices of potential end-users of Jeeves and, through their preliminary feedback on the Jeeves prototype, **describing** the ideal properties of such a tool in a real-world context.

To understand the factors influencing potential end-users' adoption of new technology in their lives, models of behaviour change have been borrowed from psychology and adapted to the technological domain. One widely used example is the Technology Acceptance Model (TAM), which conceptualises a user's attitude towards accepting a new technology as a combination of

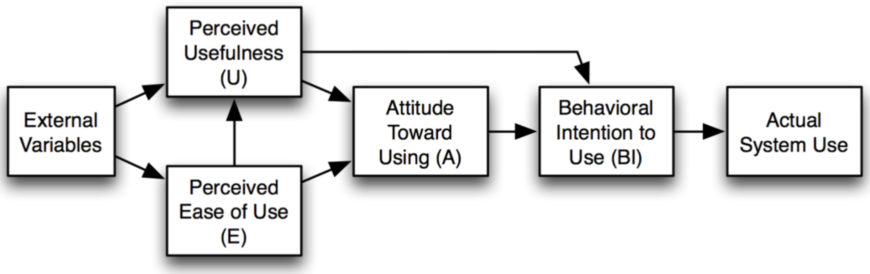


Figure 7.1
The Technology Acceptance Model, from [145]

perceived usefulness and perceived ease-of-use [145], as illustrated in Figure 7.1. Using this model as a guide to the research in this thesis, the *perceived ease-of-use* has been addressed in Chapter 6. However, given the significance of *perceived usefulness* as a prerequisite to adoption, and the paucity of published EUD research addressing this, evaluating the usefulness of a domain-specific EUD tool is an informative, and indeed necessary, contribution to EUD research in the domain of interest.

7.1 Perceived Usefulness of ESM

Chapter 6 showed that participants, with minimal instruction, can use the blocks-based paradigm to create applications considered to be potentially useful for psychology researchers or clinicians. However, a task-based usability study does not provide indication of an interface's effectiveness in its context of use. In order to fully understand the factors related to this effectiveness, more in-depth analysis is required with potential end-users, in order to identify:

- The usefulness of an ESM app in their research or practice.
- Their attitude towards end-user development of an ESM app.
- The external conditions that affect this subjective interpretation of usefulness.

Mehandjiev et al. highlight a lack of *"necessary knowledge of how to deal with problems and conflicts which are likely to emerge from the formalization of EUD"* [205, p. 372]. Much research investigates the technical properties of EUD paradigms that facilitate the transition of novice student developers from these paradigms to high-level textual programming [138]. These investigations utilise tools such as *App Inventor* and *Scratch*, where *perceived usefulness* of a tool is determined by its capability to improve comprehension of high-level programming languages, dependent neither on usage context, nor the developed applications themselves.

Although this application of EUD in education is a worthy pursuit, it disregards the implications for professionals of the formalisation of EUD in working practice. In contrast to students, professionals' perceived usefulness is highly contingent on organisational factors, as well as the capabilities of their developed apps.

7.1.1 Research methods

The ESM studies reviewed in Section 2.2 cover a wide variety of study domains and complexities, such that conceiving tasks for different domain experts is impossible. Moreover, task-based usability studies do little to predict usage behaviour in real-world contexts, and to determine design implications for incorporating EUD tools into such contexts. Therefore, a qualitative approach was taken to understand the potential applications of Jeeves, and to describe the requirements that Jeeves must satisfy for its successful adoption. In doing so, this focuses on the primary research question of *"What are the factors influencing the adoption of technology for researchers and clinicians to develop experience sampling smartphone apps?"*

7.1.1.1 Interviews

Semi-structured interviews were conducted with eight clinicians and five psychology researchers. In interviewing these two separate groups, adoption factors specific to the two different domains were obtained, both in terms of their own motivations and constraints, but also those of their participants and organisations. Interviews took place at the interviewees' location of choice, and lasted approximately 45 minutes. Focus groups were also considered for their capacity to induce discussion between participants. However, the busy schedules of both groups meant that organising a suitable time and location was impossible. Instead, individual interviews could be conducted at times that suited each professional, and ensured that each had adequate time, and assurance of confidentiality, to discuss their views openly.

7.1.1.2 Thematic Analysis

Interviews were transcribed, and qualitatively coded for thematic analysis. Coding was structured based on the research questions identified in the introduction, divided into the themes of end-users' characteristics, their attitudes towards ESM, and their requirements of an EUD platform.

Following this, a second phase of coding linked these themes to factors in existing models of technology acceptance, specifically the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) [146]. The relevant factors are *perceived usefulness* (which is further divided into *perceived potential* and *initial requirements*) and *facilitating conditions*.

Although software's *perceived usefulness* and *perceived ease-of-use* are sufficient for predicting its adoption according to the TAM, software that satisfies both factors still may not be adopted in practice. The organisational view of end-user development proposed by Mehandjiev et al. describes how the formalisation of EUD is likely to give rise to problems and conflicts, as with the introduction of any new technology into established working practices [205]. The social and organisational factors pertaining to the adoption of new technology represent the *facilitating conditions* construct of the UTAUT, which is also investigated in the following interviews.

As was discussed in the final usability study of the previous chapter, the qualitative coding of interviews in this chapter is a limitation of the work. In order to establish inter-rater reliability and strengthen the validity of the derived themes, coding should have also been performed by another researcher independently. Nevertheless, as discussed by Armstrong et al. [193], qualitative analysis of this type is naturally defined by the perceptions of the individual researcher, which are not representative of any strictly defined, objective reality.

7.2 Clinician Interviews

Although ESM is prevalent in medicine, it is most often employed in a research context, as part of a randomised control trial. However, a particularly promising application of ESM is in supporting clinical *practice*, as part of a patient-centred care model. Based on the background literature of mHealth technology in clinical practice, the interview protocol designed for clinicians was intended to elicit their thoughts on remote, patient-centred care, with regards to their current working practices. Jeeves was briefly demonstrated to clinicians to obtain preliminary feedback on its usefulness. However, clinicians were not asked to use the application at any point, thus ease-of-use factors were not addressed. Instead, discussions focused on the socio-technical and organisational barriers facing the implementation of an EUD-ESM platform in clinical practice. The interview protocol (included in Appendix C, Section C.4) was structured in three parts:

1. **Current practices** - Clinicians were asked about their current working practices, with an emphasis on dealing with patients managing chronic conditions.

2. **Use of technology** - Clinicians discussed their familiarity and experience with technology in healthcare, with a particular emphasis on mHealth and patients' use of apps for self-management.

3. **Preliminary feedback** - Clinicians were given a brief overview of Jeeves, and asked for preliminary feedback on its potential usefulness

Table 7.1
Potential benefits and drawbacks of smartphone ESM derived from clinician interviews

Stakeholder	**Perceived Benefit**	**Associated Drawback**
Patient	Patients receive instant feedback	Patients may not understand feedback
	Patients more independent	Incapable patients at risk of being forgotten
	Patients empowered to self-monitor	Patients are diverse in self-monitoring needs
Clinician	Clinicians track patients' data	Clinicians could be overloaded with data
	Clinicians acquire more accurate data	Patients may want instant feedback
	Time saved through reduced appointments	Some communication best done face-to-face

The clinicians interviewed consisted of five GPs, a pharmacist, an ophthalmologist, and a clinical psychologist. A high diversity of roles, responsibilities, and opinions with respect to patient care were exposed in these interviews.

7.2.1 Perceived potential

This section analyses the perceived potential of Jeeves based on interview feedback. Benefits and associated drawbacks derived are summarised in Table 7.1.

7.2.1.1 Potential for clinicians

Although clinicians and patients are two distinct stakeholders of Jeeves, the potential for improved patient health outcomes would reciprocally benefit the clinicians themselves, in terms of reducing unnecessary appointments and freeing up clinic time. However, opinions regarding the effect of smartphone ESM on clinic time are varied, with some clinicians concerned that their workload would be increased by providing patients with self-monitoring technology. Although this feedback relates to smartphone ESM apps in general, Jeeves is used as a concrete example.

Jeeves could remind patients of appointments

Familiarity with mHealth technology was limited, although many clinicians were aware of systems where patients are sent SMS appointment reminders, and were positive about this potential application of Jeeves, citing poor attendance rates:

“...a reminder. ’You have an appointment tomorrow!’ Yes, that sort of thing would be fab, because the non-attendance rates are really poor and that costs the NHS an awful lot of money” (C3)

Regarding this feature, it was highlighted that some patients may already actively set reminders in their mobile devices for appointments or medication, suggesting the utility of such an application from a patient’s perspective.

Jeeves could reduce unnecessary appointments

C4, a GP, explained the benefits of a *holistic* model of care that takes patients’ psychosocial issues and lifestyle habits into account. However, time-constrained appointments, and pressure to see as many patients as possible, means that GPs often rely on quick prescriptions of medication to avoid overrunning on clinic time. A further aggravating issue is that appointment slots are often used for simple reviews:

“I certainly feel like we do epilepsy reviews face-to-face which could be done remotely. Because a lot of that is asking questions or providing advice which actually you don’t need to physically be in the same room as somebody to do that” (C4)

Thus, when face-to-face diagnoses are not required, as in epilepsy reviews for example, technology such as Jeeves could increase appointment slots for acutely unwell patients.

Jeeves could allow for more accurate data

Reducing retrospective recall bias, one of the primary arguments for utilising ESM in psychology research, is also applicable to clinical practice. Clinicians explained that patients were often poor at reporting their symptoms and summarising their experience over a number of weeks:

“the doctor sees them every two weeks so then some patients they tell them ‘how was everything during the two weeks?’, ‘Yes everything was alright’...because then time passes and patients tend to forget” (C2)

Irrespective of patients’ memory bias, the act of taking an individual reading in a clinical setting is insufficient to characterise the overall status of the patient. C8 explained how, for diabetic patients, a single “HBA1C” reading is relied upon to derive longitudinal information:

“we’re relying on one single value of this HBA1C so...you might know that the patient has not been well-controlled but you don’t know why...whereas with the app it would be really informative because you can actually say where the peaks of high sugar levels are happening” (C8)

Despite the potential benefits of repeated, real-time assessment, clinicians expressed doubts about the introduction of ESM technology, summarised as follows:

Jeeves could give patients unrealistic expectations

A number of clinicians highlighted the potential problems of facilitating communication between health professionals and patients. C5 strongly opposed this idea, suggesting that patients would feel entitled to receive immediate feedback on their submitted readings:

"the expectations of people in this day when we're all used to 24 hour bing bang bong. I sent my results in at 2 o'clock this morning. Well I'm very happy for you. I was asleep. Well why didn't I get a reply when you got into work at 8 o'clock? I was seeing patients" (C5)

Paradoxically, if patients with a *clinician*-centred attitude to their health are empowered to self-manage, this could cause additional work for clinicians in responding to these patients' concerns. Therefore, patients must accept that results will not be looked at instantly.

Jeeves could overload clinicians with information

A minority of clinicians expressed scepticism that longitudinal patient information would be of any use, a symptom of *information overload*:

"What does it achieve? We're very good at capturing information and nobody doing anything with it for years and years and years. Or not doing anything with it at all. We've got all this data. What are you doing with it? Nothing. Well why did you collect it then?" (C5)

The importance of setting aside time to look at results was seen as a negative aspect of self-monitoring. Clinicians explained that the collection of patient information would be fruitless without a professional on the other end to interpret the results.

7.2.1.2 Potential for patients

Clinicians identified benefits and disadvantages for patients themselves with regards to using the JeevesAndroid ESM app on their smartphones, described as follows.

Jeeves could provide patients with instant feedback

The primary benefit of Jeeves identified by clinicians is its potential to reassure patients on their readings or symptoms through an extension of ecological momentary interventions (EMIs). Reassurance could be an effective intervention strategy for ensuring patients maintain good health between appointments. The perceived control that patients have is strongly linked to the outcomes of their conditions:

"one of the big risks to you after you've had a major illness like a heart attack is that you worry about it and that increases your adrenaline and that puts a strain on your heart, so more likely to precipitate further heart attacks. Whereas if you can feel in control and feel that you're managing your condition you're less likely to have a heart attack" (C6)

While human contact was acknowledged as a vital component of patient reassurance, clinicians also considered that certain feedback could be incorporated into an app, responding instantly to self-reported symptoms to relieve anxiety, and its dangerous consequences:

"if we transfer that into some sort of app, human contact's very important, and we still need that, but there may be something more that we can do in terms of giving people something that allows them to feel in control and helps them to calm down" (C6)

Jeeves could engage patients in their healthcare

Engaging patients to look after their health requires catering to their agenda. C1 referenced a staged model of behaviour change in order to determine which type of intervention is best for a particular patient. For example, patients in later stages of health behaviour change could be engaged through simple self-monitoring, whereas those in earlier stages are likely to require more active intervention:

"For me to tell someone 'you need to stop smoking' or 'you need to do more exercise', that's my agenda it's not the patient's agenda. What you've got to do is find out what the patient's agenda is and then give them advice which will help them to achieve that agenda" (C1)

Tailoring to patients' individual concerns, needs and wants - a core function of Jeeves - was seen as necessary to motivate healthy behaviours.

Jeeves could support remote and disabled patients

Patients in rural locations can have difficulties in attending appointments. This is also an issue for patients with certain physical or mental health conditions, suggesting that the remote provision of results or advice would be useful:

"a lot of our patients can't actually, would struggle to get into an appointment. So, and you know some of them might not be able to drive because of the medication they're on" (C3)

While it was acknowledged that many reviews and appointments would have to be done face-to-face, clinicians identified how smartphone apps could facilitate communication when this was not possible:

"This kind of mobile app, it would really improve healthcare...patients are busier so it might be the

case that they can't come to the practice because they work offshore...or they need to commute here every day" (C8)

Thus as well as saving time for clinicians, this would also reduce the commuting burden on patients. However, the potential benefits for some patients may also present themselves as drawbacks for others, discussed as follows.

Patients may not want to self-manage

While a low perceived ease-of-use may prevent certain cohorts of patients from engaging with self-management, patients' mental models of care could also result in a low perceived *usefulness* of doing so. C4 explained how some patients from lower income backgrounds or education levels may prefer a clinician-centred, paternalistic care model:

"I work in an ex-mining community. We've got high levels of socio-economic deprivation. They're a bit less motivated or less used to being self-managed, and they're quite doctor-dependent. So I think some of it's cultural as well as to how we practise." (C4)

As such, encouraging self-management routines in patients with strongly rooted clinician-centred attitudes would require significantly more effort than simple provision of technology.

Patients may not be capable of self-managing

C5 addressed a further issue that self-management technology could be potentially damaging if patients lack knowledge to contextualise their monitored readings:

"There's no point us just giving them a box of tricks to do numbers. They've got to have the knowledge that goes with it to make sense of those numbers and know when they need to react" (C5)

If patients receive incorrect readings due to a technological fault, without the independence and experience to understand these readings, self-management could indeed cause more issues than simply booking regular appointments with their clinician.

Patients may become obsessive over self-monitoring

C1 expressed concern about the "worried well", who already impose unnecessary burden on clinicians. The *worried well* are those who do not require medical treatment, but make appointments in order to seek reassurance. In giving such patients self-monitoring technology, there was an identified risk that this could induce obsession:

"It's also worried well, that it seems we often prioritise...People can come in every week and ask you to have their hair selenium checked because they're not getting enough in their diet" (C1)

In summary, both patients and clinicians could potentially benefit from technology-assisted, patient-centred care through Jeeves. However, patients must be educated on what they are expected to independently manage, and what their expectations should be of the clinician. Furthermore, even with the knowledge and experience to manage their conditions, not all patients would perceive benefit from doing so, specifically those with a clinician-centred attitude, linked to age and economic status. However, for patients both willing and able to self-manage, it could reduce their risk of returning to hospital, or to their GP for unnecessary reassurance. Empowerment of patients to self-manage could save lives and significantly reduce the burden on the NHS, and Jeeves is one way to introduce this empowerment.

7.2.2 Initial requirements

The perceived usefulness of Jeeves is not only contingent on its theoretical potential benefits, but on its existing functionality that would enable these benefits to be realised. Clinicians were asked about the context of patient management, experience with existing technology in their practice and the usability issues that they encounter. After being briefly shown the capabilities of Jeeves, they were asked who its potential end-users would be, and what facilities would need to be in place to ensure streamlined use of the software across its stakeholders.

7.2.2.1 Requirements of clinicians

Clinicians' previous experience of healthcare applications was mixed; some clinicians had been directly involved with evaluation of healthcare apps, while others had used apps for their own benefit, but not targeted towards patients. None of the clinicians had any prior experience of software development. Having determined benefits and barriers of ESM, the latter part of interviews aimed to determine clinicians' requirements of an EUD tool for creating ESM apps, which are summarised in Table 7.2.

Jeeves should make it easy for clinicians to communicate and collaborate with each other

Chronic disease patient care is a collaborative effort between GPs and specialist practice nurses. Clinicians explained how individual diseases were dealt with by nurses who had received specific training for management of that disease. Patients may also see different doctors, such that it is important to enable collaboration and understanding between clinicians. Current technology does not provide a simple means of doing so, as expressed by C3, a clinical psychologist:

"if we make a change to patients' medication, the quickest way to communicate that to the GP is to fax them, would you believe. So, if we email the GP, we have no idea when the GP might pick that up" (C3)

Table 7.2
EUD requirements for clinicians, as derived from interview feedback in Section 7.2.2

EUD Support	Barrier Addressed	Jeeves Functionality Required
Healthcare staff collaboration	Patients have their conditions managed by many staff with different specialities	Community support approaches, including shared editing and semantic annotation
Personalised Reminders	Dependent, incapable patients are at risk of being forgotten	Prompts delivered at dates and times personalised to patient preferences
Automated Feedback	Patients need reassurance that their readings are being acted upon	Advice prompts or follow-up surveys delivered conditionally on survey responses
Individual Tailoring	Patients are at different stages and require protocols tailored to them	Protocols that deliver different messages or surveys conditional on patient preferences

It also emerged that comorbidities would need to be taken into account. Patients with a chronic disease such as diabetes often have a coexisting disease such as asthma or hypertension. While one nurse would be capable of authoring an action plan for managing a patient's diabetes, this would require input from an asthma nurse if such a comorbidity was present.

Jeeves should support personalised reminders

Clinicians were doubtful that certain patient cohorts would embrace self-monitoring. Some are content with seeking advice and reassurance from their GP, such that this dependence would prevent engagement in proactive health behaviours. Indeed, such patients are often incapable of self-monitoring, and acting appropriately on received feedback.

However, patients less likely to engage in self-monitoring are those for whom it would be of greatest benefit - they are often elderly, managing comorbidities, reliant on various medications to be taken at different times, and also reliant on regular appointments with their GP. Thus, the ability for clinicians to implement tailored reminders for appointments and medication into a patient's app was perceived as a significant advantage to an EUD approach. In this case, the ESM functionality would not necessarily incorporate assessment.

Jeeves should allow clinicians to implement automated feedback

In addition to the barriers presented by passive or uninformed patients, clinicians identified barriers imposed by those who are expectant of instant feedback, compounded by an overload

Table 7.3
Ease-of-use requirements for patients, from interview feedback in Section 7.2.2

App Requirement	For whom	JeevesAndroid Functionality
Accessibility	Elderly patients with eyesight and coordination issues	Ability to adjust size of text and buttons, ability to send and receive audio data
Integration	Proactive patients engaged in self-monitoring	Ability to interface with current self-monitoring technology such as blood pressure monitors
Flexibility	Passive patients with regular medication/appointments	Ability to easily adjust dates and times for appointments or taking medication

of incoming patient information. As a resolution, clinicians valued the possibility of adding in automated responses to patients monitoring their data. Through automated feedback, clinicians appreciated the possibility to provide reassurance or advice to patients on their readings, while minimising their own workload. Such feedback could also prompt patients to make an appointment with a clinician, thereby increasing the safety of high-risk patients.

Jeeves should allow tailoring to individual patients

A patient-centred care approach relies on the incorporation of patients' own goals, motivations and constraints, which determine whether or not they are likely to engage in treatment. C6 described these factors as the patient's "agenda":

"Healthcare people, not just doctors, have agendas, but the patients often come in with a totally different agenda, so it's about sorting out what the patient's agenda is and how you can use that to improve their physical wellbeing." (C8)

Hence, if EUD-ESM is to be integrated into clinical management of chronic care, it is therefore necessary to provide simple tailoring to individuals.

In summary, corresponding drawbacks to previously identified benefits of Jeeves are those that must be alleviated with minimal effort by clinicians. EUD support for collaboration, implementing timely reminders, implementing automated feedback, and individual tailoring, could address these issues.

7.2.2.2 Requirements of patients

Clinicians were concerned about the usability of a smartphone app by many patients managing chronic disease. In addition to requirements for the Jeeves visual programming environment itself, clinicians described requirements for developed apps, summarised in Table 7.3.

JeevesAndroid should be flexible to patients' needs

For patients who would not directly engage in self-monitoring, it should be simple to set up reminders for different aspects of their healthcare. C3, a clinical psychologist, mentioned that either patients themselves, or their carers, would add appointment reminders on their mobile devices.

"Interestingly quite a few of our patients say they've set reminders on their phone for things like daily reminders of it's time to take their medication. So they've actually done that themselves, or somebody's usually done it for them to be fair" (C3)

Additionally, many of these patients are required to take medication at particular times in their daily routine. C8 suggested that it should be possible for patients to input their waking and sleeping times, as well as regular mealtimes, to ensure that medication is taken.

JeevesAndroid should be accessible to the elderly

Clinicians acknowledged the growing acceptance of smartphone technology in elderly populations. It is particularly desirable that apps are accessible to such populations, in which chronic disease is most prevalent. The physical constraints of elderly patients must be taken into consideration, as explained by C7:

"Practical hand-eye coordination, hand skills, if you've got quite advanced arthritis it's very difficult to handle a mobile phone for apps, I suppose lots of old people struggle to tip-tap on the screens and work out what's going on" (C7)

As a potential solution, larger button sizes, verbal control, and minimalist user interface design were suggested as useful enabling features for elderly users.

JeevesAndroid should support patients' current self-monitoring

Clinicians also expressed positive opinions with respect to the feasibility of self-monitoring, given that many patients are already engaged in doing so. It emerged that patients with chronic diseases are often highly skilled in their management, and willingly purchase blood pressure or glucose monitors to facilitate their independence.

In order for the JeevesAndroid app to become an asset to their healthcare, rather than a burden, it should be possible to integrate with patients' current self-monitoring practices. It was suggested that Bluetooth and other wireless data transfer technology could enable patients' results to be seamlessly uploaded from their existing devices.

7.2.2.3 Summary

In summary, the needs of particular patient groups are a key concern in clinicians' adoption of technology designed to support them. For example, clinicians would not consider a smartphone self-monitoring app for elderly patients managing complex conditions, but were able to conceive alternative applications of Jeeves for such patients if the flexibility to implement personalised reminders was available. The constraints imposed by clinicians' time, and target patients' familiarity, suggests that adoption of such technology in clinical practice would be at a small scale, potentially growing as the aging population becomes more adept with technology.

7.2.3 Facilitating conditions

The *facilitating conditions* construct, used in the UTAUT model of technology acceptance, represents *"the degree to which an individual believes that an organisational and technical infrastructure exists to support use of the system"* [146]. As with other constructs, facilitating conditions are dependent on the technology's usage context.

In the context of clinical practice, constraints imposed by the NHS present significant barriers to EUD adoption. Healthcare transformation is difficult in a system where it must take place at a national level. Thus, facilitating conditions must be fulfilled to enable the transition from clinicians' *intention to use* to their active *usage behaviour*. These conditions, with their relationship to each other and relevant TAM constructs, are illustrated in Figure 7.2.

7.2.3.1 Condition D: Patient data must be secure

One organisational barrier identified by many clinicians was the difficulty of accessing patient information, which is restricted outside of the NHS-certified computer network. Security issues would therefore prohibit remote access to this information. The interviews were conducted approximately six months after a major security breach of the NHS computer system, such that the transfer of patient information was a particularly sensitive issue for clinical practice.

"I think we wouldn't be allowed to, that's the difficulty. We would not be allowed to put that on a mobile phone because it's got patient data. That's the issue" (C3)

Thus, a prerequisite of integration of Jeeves into the NHS is assurance of security. Without the guarantee that sensitive patient information will not be accessible with malicious intent, data transfer between mobile devices and NHS computer systems will not be possible, thus eliminating the benefits of real-time monitoring.

7.2.3.2 Condition B: Clinicians must trust the resilience of developed apps

From a technical perspective, clinicians were wary about the reliance on technology to capture abnormalities and give patients immediate feedback at critical times. Quality assurance of apps is critical in a medical context, where the health and wellbeing of patients could be put at risk. Clinicians had concerns about what would happen if such an app were to malfunction:

"if suddenly you don't have network coverage, then your reading might not get to the platform or whatever or the server in two hours, and then if it's a high reading...you might not be able to receive the message soon enough" (C8)

While the offline functionality and automated feedback capabilities of Jeeves alleviate some of these concerns, reliability remains a major barrier for acceptance not just by individual clinicians, but the health service to which they belong.

7.2.3.3 Condition C: Jeeves must undergo trial evaluation before adoption into the NHS

Data security and app reliability are essential facilitating conditions for acceptance of EUD technology into clinical practice. However, in order to formalise this assurance of app quality, these conditions must be satisfied by an NHS trial evaluation in order to ensure that it is ethically sound and will do no harm to patients.

"You'd need to set up some kind of evaluation which would require a trial, require NHS ethics...that is a major obstacle to the embedding of this into NHS care, it requires an official scientific, credible, systematic evaluation to show that it doesn't do any harm" (C6)

7.2.3.4 Condition A: Clinicians must be allowed time to get used to Jeeves

Although time was a critical factor across all clinicians, GPs in particular reported a lack of time given their high workload. Ironically, clinicians would struggle to adopt time-saving technology because the initial time to introduce it would be too costly, as expressed by C4:

"...I know that takes like a minute or two, it's a minute or two I might not have. The times are very very precious...this will free up time ultimately. But it's that initial jump, isn't it?" (C4)

GPs expressed frustration that, paradoxically, they need more time to discuss empowerment and health promotion that would reduce unnecessary appointments. However, this high number of appointments forces them to limit individual patient time to 10 minutes.

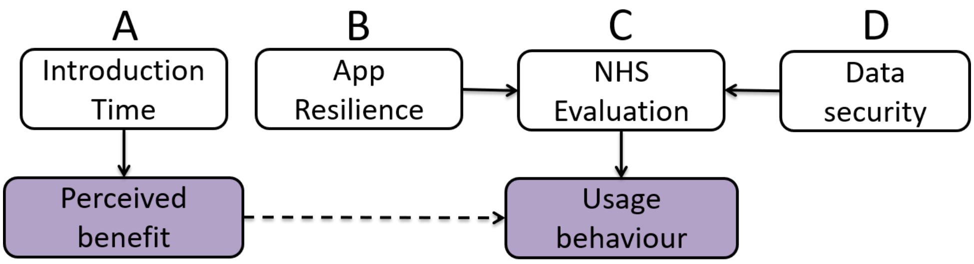


Figure 7.2
Facilitating conditions for clinicians (white) and how they relate to TAM constructs (purple)

7.2.4 Summary

Interviews with clinicians uncovered various considerations for the integration of ESM into clinical practice. While the potential benefits for clinicians and patients alike were recognised, these were contrasted with potential drawbacks that would preclude its adoption with some patient groups. The perceived usefulness and ease-of-use for both clinicians and patients are necessary factors for adoption, notwithstanding barriers imposed by the NHS. While ease-of-use was not directly addressed, one clinician was explicit about the perceived benefits he saw of the blocks-based approach:

"I think the idea of being able to drag and drop buttons and to have obviously kind of computer language behind them which is programmed for those buttons to be active and work is an incredibly powerful and slick way for a person who's naive to the language to be able to generate something that's going to work, and I will say aloud quite specific, tailoring to the local population, for whatever their cultural needs and demands are" (C7)

This was particularly positive feedback, given that clinicians were not asked specifically about the design of the visual language.

7.3 Researcher Interviews

Psychology research is the primary domain in which ESM is utilised. The methodology was initially developed as a means of obtaining ecologically valid research data, whereas its application to patient healthcare in clinical practice is comparatively novel. As such, it was expected that there would be greater enthusiasm and envisioned applications of Jeeves within the research community. The interview guide (included in Appendix C, Section C.4) was similar to that of clinicians' interviews, with questions adjusted to reflect working practices in research:

1. **Current practices** - Researchers discussed how they currently collect participant data, with

the barriers and difficulties encountered in doing so, and their motivation for conducting longitudinal data collection.

2. **Use of technology** - Researchers described the technology used in their data collection and analysis. The learning curve of referenced technologies, their difficulties and disadvantages, and factors motivating their use were covered in this section of the interview.

3. **Preliminary feedback** - As with clinician interviews, researchers were provided with a brief overview of the Jeeves platform, and were invited to give feedback on the various features of the interface.

Interviewees were all social psychologists, with an interest in the contextual factors of human behaviours and attitudes. While it was only possible to interview five psychologists, their research areas, and characteristics of their participants, were sufficiently diverse to obtain a range of considerations for adoption of Jeeves. Broad themes of research were gender differences, emotional disorders, social identity, crowd behaviour, and intellectual disabilities.

7.3.1 Perceived potential

Table 7.4 summarises the potential benefits of Jeeves to psychology researchers, and their participants. Given that the applicability of longitudinal methods are well-established in psychology research, it was unsurprising that interviewed researchers perceived potential benefits as stronger than their corresponding drawbacks.

Unlike in the clinical domain, where patients and clinicians could benefit equally, the advantages of ESM in research are largely specific to researchers themselves. Study participants often have little intrinsic motivation to engage with a study, instead receiving financial compensation for the burden imposed by repeated assessment. However, there are exceptions to this, particularly in working with sensitive participant groups, such that Table 7.4 shows how psychological benefits could impact on participants themselves.

7.3.1.1 Potential for researchers

Many of the perceived potential benefits of Jeeves related to researchers' difficulty in conducting ecologically valid research.

Jeeves could enable compliance monitoring

As discussed in Chapter 2, compliance remains a major issue for studies where participants are required to maintain active participation outside a lab environment. P4 explained how tracking

Table 7.4
Potential benefits and drawbacks of Jeeves for researchers, derived from interview feedback in Section 7.3.1

Stakeholder	Perceived Benefit	Associated Drawback
Researcher	Researchers can monitor compliance	N/A
	Researchers acquire in-the-moment data	App becomes disruptive
	Researchers acquire contextual info	Participants become suspicious
	Time saved through remote research	Some research best done in the lab
Participant	Participants engage in self-monitoring	Participants not capable of doing so
	N/A	Research distresses participant

the number of completed surveys, as well as the time taken to complete these surveys, could be used to motivate participants with additional financial compensation:

"if you could have a mechanism I guess of...you know recording if you complete all these bits we'll put you in a prize draw...and you can see if they've done the whole survey in 10 seconds" (P4)

The ability to automatically, or indeed manually prompt participants to complete surveys was valued by P3, who explained his use of *Qualtrics* software in order to manage participant compliance, suggesting that this would be a useful feature:

"[Qualtrics] keeps track of who's not responded yet so you can send up a follow-up email to only those who've not responded...allows you to interact with your participant pool" (P3)

Jeeves could eliminate recall inaccuracy

P3 explained how participants saw cross-sectional surveys as a distraction or intrusion to the event of interest. Consistent with literature reviewed in Chapter 2, researchers mentioned the disadvantage of retrospective recall, where a question regarding an event of interest is too far removed from the event itself. Experimental methods also precluded in-the-moment perception of events, introducing recall and contextual bias:

"One task, which was to describe how they would solve hypothetical problems and I said what

memories come to mind, so it's not the same thing as being in an environment where you might have an argument with somebody and then, what's really going through their mind" (P5)

Alleviating recall bias and ecological validity issues caused by lab-based data collection was the most significant perceived benefit. Researchers described how a longitudinal experience sampling study would help to alleviate the memory biases that occur due to the time lapse between an event and its reporting.

"The closer you can get to the actual event, and treating each event as a unit, as opposed to people trying to kind of impose their expectation about how those things should be experienced, we get a bit closer to the raw experience itself in some way." (P3)

Jeeves could enable capture of contextual information

Time, location, and user-triggered assessments were all perceived as important in practical application of Jeeves. While clinicians were less enthused about the concept of prompting patients in different contexts, researchers were able to perceive a range of reasons for doing so. P5 explained a study she hoped to conduct investigating contextual influences on participants' mindfulness, such that location-based triggers would be of use:

"'location outside', how are you feeling now, that would be useful. That's something we want to develop. 'Where are you? Are you in your bedroom, meditating? Are you outside?' That could be extremely useful" (P5)

As well as triggering assessments based on location, researchers perceived value in simply capturing location information at particular time points to supplement self-report. Allowing participants to manually provide additional context such as images or video recordings was also suggested as a rich information source:

"maybe...'just send us a photograph of what's going on right now, in the scene, at this moment in time' especially if that was then location and time stamped that could be quite good" (P4)

Jeeves could save time over traditional data collection

Researchers use a variety of data acquisition methods, including paper handouts, online surveys, lab-based experiments, and face-to-face interviews. Their particular areas of research determine the type of methods that are employed. For example, P2 works with intellectually disabled participants, so often conducts interviews due to difficulty in comprehending written materials.

The cross-sectional nature of most research was apparent, and researchers were explicit about the disadvantages of this approach. P4 is interested in experiences of participants at crowd events, often visiting these events of interest and handing out paper surveys for data collection.

Experimental lab research also poses disadvantages, including time and recruitment issues:

"it's difficult to get people into the lab in the first place, to recruit them, it takes a lot of time to organise 'cause you can only do 5 or 6 a day at most, and then once they've done that study they can't really do another one" (P4)

Regarding this issue, P2 discussed the benefits of using mobile methods, particularly in gathering data in difficult settings, and eliminating the need to manually transcribe data from paper surveys:

"you can collect data in the field far more easily, collect data in a variety of settings. You don't need a desk to sit and write, you don't need paperwork to collect" (P2)

These benefits are not without their challenges, however, and salient issues were identified by researchers that show how realistic expectations of Jeeves must be maintained.

Jeeves could provoke non-compliance

While compliance monitoring is possible through real-time data transfer, experience sampling itself could negatively affect participants' compliance and honesty if not used appropriately. P2 explained that repeatedly asking the same questions to intellectually disabled participants could cause reporting inaccuracy:

"You would get participants who would question why you were asking them the same thing again, because they'd already answered that question. Didn't you ask them that two weeks ago or didn't you ask them that 4 months ago?" (P2)

Jeeves introduces a lack of control

Experimental rigour and control are integral in much of the work that interviewed researchers conduct. Thus, they expressed concern about the lack of control that they would have on data collected with Jeeves:

"There are downsides of people completing measures online because you don't have any control of how they're doing it...listening to blaring music while they're doing it, you know you don't have control over the environment" (P5)

Particularly in experimental work, researchers expressed a need to measure millisecond reaction times, and had found this to be a disadvantage in their application of *Qualtrics* software:

"([Qualtrics is] not very good at measuring the timing of keystrokes, button presses, whatever. It does it, but not brilliantly...The design of the thing is not really set up for experiments" (P1)

In social psychology, a large proportion of research is conducted experimentally within a lab setting. While Jeeves facilitates the setup and management of field research, its limited functionality in relation to lab experiments would have to be accepted by researchers.

7.3.1.2 Potential for participants

Regarding the potential benefits to participants, P2 described how applications of Jeeves could be *proactive* or *reactive*. Proactive uses are those where Jeeves could motivate lifestyle changes and strategies without the need for external stimuli, such as encouraging participants to self-monitor. Conversely, reactive uses are those where participants would engage with Jeeves in response to an external event, such as a stressful situation, in order to seek help or reassurance. Three of the five researchers either investigate particularly sensitive issues, or work with mentally ill participants. In each case, participant well-being and confidentiality has to be addressed.

Jeeves could support self-monitoring

Similar to a benefit of Jeeves in clinical practice, P2 suggested that with intellectually disabled participants, Jeeves could be used as a monitoring tool to self-regulate and record dietary habits. The possibility for participants to view and track their data over a period of time, and to receive automatically prompted feedback from an app, could support them to independently manage their health and well-being. P2 additionally described how mobile phone use in intellectually disabled populations has dramatically increased in recent years, and that such technology is seen as an asset to their independence. Self-monitoring could thus be linked to a carer or alternative emergency contact, thereby providing benefits in practical applications to participants, rather than solely research-based advantages:

"...if you'd something like this with a button that alerted carers, that's better, that's less intrusive than walking about with a band...for self-monitoring, bullying purposes you've got some sort of button where they can communicate with people and say 'I'm not feeling safe' " (P2).

7.3.2 Initial requirements

Issues regarding software were discussed in far more detail with researchers than with clinicians. While the latter group are constrained to NHS-approved patient management software, researchers use a wider range of technology, and similarly have differing motivations and experience with learning how to use new applications.

The learning curve of researchers' technologies of choice was probed in order to find what level of effort was considered acceptable, and what support they found useful during the learning process. It emerged that the usability in performing complex operations was not considered as important as the overall utility of the software. However, a low entry threshold for performing simple tasks

Table 7.5
EUD requirements for researchers, as derived from feedback in Section 7.3.2

EUD Support	Barrier addressed	Jeeves Functionality Required
#1: Researcher Collaboration	Study protocols must be understood between research groups	Community support approaches, including shared editing and semantic annotation
#2: Peer-oriented support	Software without adequate tech support will not be adopted	Configuration files, semantic annotations and community support forum
#3: Low barrier to entry	Unclear options and resistance to change can affect usability	Incremental development support with low viscosity and high visibility
#4: Meta-design framework	Researchers won't adopt new software if their needs are not met	Evolutionary system with social infrastructure for obtaining new requirements
#5: Contextual triggers	Apps that trigger at inappropriate times are disruptive	Triggers based on participants' location, or within times tailored by participant
#6: Automated assurance and support	Sensitive data collection may upset participants	Assurance prompts delivered on survey responses and support resources on request

was considered to be necessary for adoption. Two of researchers' initial requirements are to enable potential barriers identified in Table 7.4 to be easily addressed, and the remainder are related to the learning and use of the environment itself. These requirements are summarised in Table 7.5, and are numbered in relation to the sections that describe them in more detail.

7.3.2.1 #1: Jeeves should support collaboration within and between research groups

In collaboration with peers, or as supervisors to students, researchers work in teams with varying experience, where community support could improve ease-of-use [116]. P1 expressed the benefits of peer support and working with examples, when learning *SPSS*:

"That was so useful just to be immersed in an environment with somebody I could ask questions...I came away with a memory stick with a whole load of code on it that I had played with already...so whenever I'm trying to do something new I can think, oh, it's a bit like that example" (P1)

P2 explained how research is typically conducted as part of a team with different specialities.

Often, the researchers most knowledgeable about study requirements lack the technical skill to implement these requirements themselves.

"You've usually got somebody who's very up on the evidence, very up on the research, but not necessarily technically...that competent...then you've got somebody else on the team that says right I know how to do [programming]...a minimum of two people" (P2)

Finally, two researchers mentioned the "replicability crisis" in science, and how research groups developing software may therefore be indirectly developing for future groups to ensure replicability of studies. P5 explained how the crisis is partly due to poor reporting of the methodology applied:

"it's extremely important when you write up your results that you describe what you've done in your methods, write clearly and precisely so other people can replicate it" (P5)

With respect to this, the concrete representation of a study specification employed in Jeeves could be valuable for knowledge transfer within and between research groups.

7.3.2.2 #2: Jeeves should scaffold learning through peer-oriented support

Acceptance of technology is contingent on adequate support for learning and applying its features. P1, in transitioning from *SPSS* to *R* statistical software, expressed how such support had enabled her to learn complex functionality. She consults documentation when performing complex tasks, rather than learning how to do these tasks independently:

"By the time I started using it there was that critical mass of people who were developing wikis and stack overflow and this that and the other...I'm not very good at using R but I am good at Googling how to do what I want to do" (P1)

While ease-of-use was considered beneficial, P2 explained how researchers are often not those implementing the software, for which they instead seek external support:

"there's a lot of platforms that we use for creating data where we will go for technical support to implement the research idea...it's more common for non-specialists to be coming and saying 'could you set that up for me' " (P2)

Peer support appears to be salient in lowering the learning curve of software. If this is not available, researchers will readily engage with higher-level technical support for accomplishing their specific development tasks. Thus, as a particular design consideration, an EUD tool should be designed for a spectrum of end-users. At one end are novices, who consult documentation and relevant examples to develop a study fit for their purposes; at the other end are "power users" who would explore all features of the interface, and scaffold typical examples that novice end-users could apply.

7.3.2.3 #3: Jeeves must allow researchers to perform simple tasks

The various functions in Jeeves for tailoring apps to individuals and triggering based on contextual information were perceived as useful to researchers, but could reduce usability by introducing unnecessary complexity to novice users. Parallel to the needs of clinicians, researchers valued the idea of pre-created examples with standardised questions, that could then be tailored if necessary. For example, P2 appreciated the ability to tailor based on attributes, but suggested that this functionality could be introduced in time:

"It's a question of...is there a point at which you need to introduce attributes or do you need to have that there from the start? I don't know what the answer to that is. You would always start with survey design or blocks, whichever" (P2)

Distinct levels of technical self-efficacy arose within the small sample of psychology researchers. This is particularly salient in interview excerpts from P3, an experienced software developer, and P1, with little programming experience. Both described accomplishing identical tasks, with their contrasting experiences highlighted in Table 7.6. While both researchers felt capable of developing the same conditional survey behaviour, the inexperienced user found a less sophisticated workaround to do so. While complex functionality should be provided, if it is also commonly used functionality, then simplified workarounds must be enabled for non-expert users.

7.3.2.4 #4: Jeeves must allow experienced researchers to perform more complex tasks

The desire for both simplicity, as well as complexity when required, was expressed by all researchers. P1 expressed how the limits of what she could accomplish with *SPSS* forced a complete transition to the more complex *R* software:

"SPSS is very easy to pick up, but you reach a point very quickly where what you want to do is beyond the scope of what it really does and then you have to give up and move to R and start at the bottom of the learning curve again" (P1)

Moreover, researchers were averse to software that was inflexible to their diverse needs. P3, a researcher with a technical background, resorted to manual development of generic software to accomplish his goal that purpose-built software did not provide:

"In a sense what I'm doing is press-ganging a more generic piece of software into this kind of mechanism. I mean it can be done, but it's sub-optimal in that sense" (P3)

From this perspective, it is important that non-programmer researchers are able to request functionality that meets their needs, otherwise this functionality will be sought in more complex software, or in a professional programmer.

7.3.2.5 #5: Jeeves should support contextual triggers

In situ, repeated self-reports were acknowledged as potentially disruptive. However, researchers suggested that contextual triggers could minimise the number of unnecessary interruptions. P4 described a potential application of the location trigger functionality to minimise interruptibility:

"If you had the location, you could, as soon as they left...some section of the pilgrimage, at that point it's appropriate right now to ask them what happened...it's going to be fresh in their mind without interrupting their experience, that'd be great" (P4)

Event-contingent sampling functionality was considered necessary for participants to identify their own contexts of interest, which other time-based sampling strategies could potentially miss:

"For the participant to say 'ah, something has just happened I want to tell you about it now' rather than necessarily waiting for me to bug them...having that flexibility, it looks very useful" (P1)

In general, researchers supported a variety of triggers including those contingent on implicit and explicit context, capturing required information with minimal disruption to participants' lives.

7.3.2.6 #6: Jeeves should support automated reassurance

Researchers, similarly to clinicians, supported the potential for delivery of feedback to participants, both automated and person-to-person. For example, P2 described working with intellectually disabled individuals, who have a carer or guardian whom they contact for support. Incorporating a means for direct communication between participants and carers was thus considered useful in non-research applications of Jeeves.

As well as enabling such support from human sources, this could be supplemented with automated support for participants for whom direct researcher contact may not be feasible. Both researchers and clinicians endorsed this possibility:

"Wherever we do research where there is a possibility of causing distress we have to take that incredibly seriously...we could automate provision of support to some extent, or at least automate the beginnings of providing support" (P1)

P5 works with patients with clinical depression and explained that participants could react negatively to certain feedback, such that provision of resources personalised to the feedback

Table 7.6
Two researchers discuss the "piping" functionality of *Qualtrics*

Novice user (P1)	Power user (P3)
"you want to assign someone to a condition then pipe them down one route or another...there IS a way of doing it neatly, but I STILL don't know what it is, and it's much easier to tell my students **look, just duplicate everything and then have a thing at the beginning where you push people towards one of these things**"	"most people don't know about that stuff so for example if they want to give people three different versions of a question **they end up doing three different surveys and then they kind of somehow randomly allocate people to three different URLs**...rather than using the facilities that allows them to keep all that data in one place"

given would be a useful feature:

"...if it was negative feedback for whatever reason...if you were giving them scores on a depression inventory, and you had appropriate resources that you could attach to the feedback" (P5)

Practical considerations regarding ease-of-use were centred on participants' willingness to comply with a study. Researchers expressed concerns that an app that is difficult to use, or an app that is intrusive regarding the data it collects, will be quickly removed by participants. Ease-of-use requirements for participants are summarised in Table 7.7.

7.3.2.7 JeevesAndroid should be non-intrusive in participants' lives

To varying extents, all forms of in situ research intrude on participants' everyday lives. P4 explained that getting crowd members to fill in paper surveys was difficult:

"They don't want to do paperwork you know....'I'm not interested in that, I'll do your study but I don't wanna read this 3 page document and I don't really wanna start signing stuff' " (P4)

P3 additionally explained his research involving professionals coping with stressful situations. Such participants were non-compliant due to the time commitment of filling in a long, cross-sectional survey:

"There's a couple of problems with cross-sectional studies...They're very busy people and so getting them to set aside half an hour, to actually do it...is actually really difficult" (P3)

Thus from a compliance perspective, an experience sampling app should be minimally disruptive. Ideally, participants should not be prompted at inappropriate times, and apps should be minimally

Table 7.7
App requirements for psychology research study participants

App Requirement	For whom	JeevesAndroid Functionality
Accessibility	Participants with learning difficulties or language barriers	Ability to communicate using symbols and emojis, and different languages
Integration	Participants with different mobile devices, and existing self-monitoring technology	Ability to interface with current self-monitoring technology, and compatibility with iOS devices
Assurance	All participants, particularly those providing sensitive personal information	Ability to provide assurances to participants, professional appearance and smooth operation

burdensome to interact with, to ensure that participants do not uninstall out of frustration:

"The more user-friendly you can make it the better...you get some apps that are just clunky and they don't really work and you're just thinking 'I can't be dealing with this' " (P4)

Integration with existing technology, a requirement suggested by clinicians for proactively self-monitoring patients, was also suggested to be an asset for research. Physiological data from commercial exercise bands was seen as potentially useful data, and would alleviate participants from the burden of having to enter this data themselves.

Additionally, researchers were enthusiastic that an app should be compatible with as many devices as possible, thereby increasing the number of potential participants:

"You don't want some software or app that's only going to work on high-end machines because most people aren't necessarily going to have those" (P4)

7.3.2.8 JeevesAndroid should assure participants of anonymity and confidentiality

Researchers expressed how conducting remote research requires participants to feel assured that their data is being collected with complete confidentiality. The primary means of doing this is by explicitly giving participants assurances throughout a study:

"...making it very clear to them what was involved in terms of sharing of information...it wouldn't be a sophisticated process but it would need to be something that is done very clearly" (P3)

However, assuring participants of sensitive data storage also relies on quality of the app's appearance, such that there is an explicit need for an app to look professional to the participant:

"It needs to look professional, and intuitive...if it's really slow, going between pages, then people are gonna give up. People generally have quite a low threshold I think for some of the stuff" (P4)

7.3.2.9 JeevesAndroid should support interaction with participants with different skills

Interestingly, the need for a wide range of accessibility options was equally as significant in research as in clinical practice. Populations with mental and physical disabilities are often ideal participants for social psychology research. Accessibility thus refers not just to smartphone compatibility, but also the compatibility constraints of the participants themselves. For example, P4 expressed an interest in extending the remote research capabilities of Jeeves to conduct research in different countries:

"You could communicate to them in a language of their choice, especially if you could time...per location so when you get to this crossroads you get this pinged up in your preferred language" (P4)

Irrespective of language, it was also acknowledged that many participants would have poor literacy, such that other means of communicating information would need to be used. P2 suggested graphical depictions of instructions and answers to survey questions:

"Who is it suitable for? If you try and make it too text-heavy you're talking about a fairly narrow group of people but if you open it up with emojis and symbols then you've got something that's more user-friendly" (P2)

7.3.3 Facilitating conditions

Organisational barriers that could potentially impose upon researchers are far less prevalent than those experienced by clinicians. Researchers are free to use any technology for participant study and management, conditional on its approval by the university ethics committee. This approval process has a turnaround approval time of approximately two weeks at most, in stark contrast to clinical practice, where the rigorous NHS ethical approval process could take many months. Nevertheless, certain external conditions must be met that influence both researchers' perceived benefit, as well as their eventual usage behaviour. As before, the conditions are labelled, and their relationship to TAM constructs illustrated in Figure 7.3.

7.3.3.1 Condition B: Jeeves must be affordable by individual researchers

Researchers' motivation to use a technology was largely dominated by its cost, in particular reference to *Qualtrics*, which was used by all five interviewed researchers. It emerged that

Qualtrics is free and easy to download, as the university has a site licence, such that this is a prominent reason for its widespread use.

"Affordability is obviously a big thing so...one of the reasons we were speaking about Qualtrics because not only does it seem to be the market leader but it's also...we have a university licence for that which is a major, a major issue for using that" (P4)

Contrary to the original Technology Acceptance Model, usability is a lesser concern than organisational factors, in particular affordability and accessibility. P3 explained that he would be more willing to use *"a good cheap option that requires a little bit of threading to make it work"* as opposed to a more expensive, purpose-built solution. Researchers are willing to work with sub-optimal technology, either sacrificing the sophisticated capabilities, or engaging in workarounds to get these capabilities.

7.3.3.2 Condition D: Participant data must be secure

A further concern shared by clinicians and researchers is ensuring the security and confidentiality of participants' data. Social psychology research often involves the storage and transfer of sensitive participant information, such that security is paramount. Research conducted within the university undergoes an ethical approval process, where the method of data acquisition, and its subsequent storage, must be described in detail. Scientific journals also require assurance of the security of the methods used:

"In terms of reassuring not only our institutional methods committee but publishers now, so for example, PLOS One. I'd need to be able to tell them this is the kind of encryption, this is what's happening" (P1)

7.3.3.3 Condition A: Jeeves should integrate with existing research technology

Researchers were enthusiastic about the ability to integrate with existing technology. Outside of the functionality and usefulness of Jeeves itself, researchers expressed the importance of maximising compatibility with other widely used research software, including *Qualtrics*:

"I can certainly see some advantages in the kinds of things that it does really well. Inevitably there's gonna be things it can't do...having that capacity to have that inter-operability, plug-in capability, developing sort of thing would be great" (P3)

It is also important for researchers to have participant data in a format compatible with analysis software such as the aforementioned *SPSS* and *R*. Unlike clinicians, who in practice would use Jeeves to monitor an individual patient, researchers place more emphasis on the analysis of

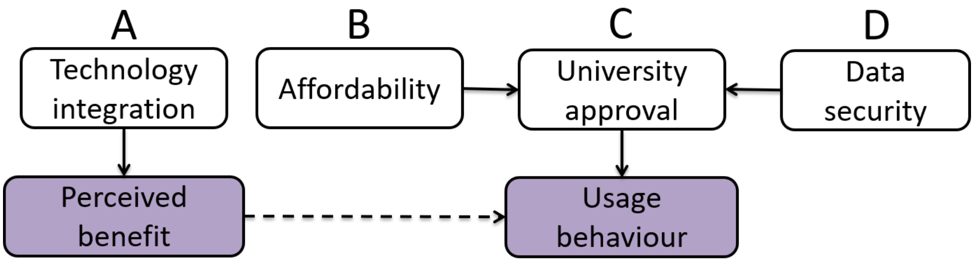


Figure 7.3
Facilitating conditions for researchers (white) and how they relate to TAM constructs (purple)

aggregated results from multiple participants.

Similarly, integration with *Prolific*, software that researchers use for participant recruitment, was seen as highly desirable. One of researchers' key motivations for conducting online research is the accessibility to a much larger participant cohort. The limitation of small sample sizes was cited by researchers as a contributing factor to the replicability crisis:

"Now in social psychology we've had something recently called the replicability crisis...one argument is that our sample size has not been enough" (P4)

Condition C (University approval) is included in Figure 7.3 as it relates to the condition of NHS approval in clinical practice. However, as explained, organisational approval is less of an issue in a university setting, and as such did not merit significant discussion in interviews.

7.3.4 Summary

Researchers were able to envision the practical application of Jeeves more clearly than clinicians. Primary concerns related to affordability and technical support; their areas of research were particularly well-suited to Jeeves, and perceived usefulness raised few conflicts. As with the clinician interviews, ease-of-use was not assessed directly, but one researcher's unprompted comment on the blocks-based structure suggested that this would be an appropriate paradigm to enable rapid understanding between different researchers:

"Again for me I think the visual thing looks like it's awesome, works really well. If obviously, you've obviously designed things to actually capture the key relational information in the smallest space as possible, so that...you can sort of see what's going on at a glance" (P3)

7.4 Discussion

The interviews of clinicians and researchers in this chapter provided a wealth of information on the factors for adoption of EUD-ESM technology in the respective domains. Having established the perceived usefulness, ease-of-use, and facilitating conditions of Jeeves in both domains, this discussion relates the findings to the first two research sub-questions defined in Chapter 1:

1. What difficulties do researchers and clinicians face that allowing them to develop experience sampling apps could alleviate?
2. What are the necessary features of apps that facilitate remote assessment of research participants or clinical patients in their natural environments?

7.4.1 End-users and their difficulties

The potential designers of ESM apps are not strictly "professional end-user developers", in that they are not typically experienced in developing software to accomplish their professional goals [113]. However, their motivations are in line with such professionals, in that EUD would be undertaken to facilitate working practices.

In the background work of this thesis, "psychology researchers" and "clinicians" were loosely defined as the stakeholders for ESM apps. The interview process enabled a clearer definition of these concepts.

Psychology researchers interested in ESM apps are social psychologists, aiming to answer questions about human behaviour. While there are numerous publications in psychology journals that utilise ESM, practical details in this literature regarding how researchers communicated, collaborated and obtained the necessary software, are not provided. Interviews gave greater insight into working practices and their associated difficulties. A summary of psychology researchers follows:

Psychology researchers

- They are not typically programmers, and may rely on others to develop software
- They typically work in teams of varying knowledge and technical skill
- They rely heavily on software that is freely available to them
- They are willing to invest time in learning software provided that the support is available to them
- They have difficulty recruiting participants and appreciate online data collection
- They are interested in the everyday experiences of participants but are faced with issues of intrusion, compliance and recall bias
- They work with participants who are often highly sensitive or intellectually challenged

Further, it emerged from interviews that although "clinician" is a broad term, encompassing a number of individuals with varying degrees of time and knowledge, it is not general enough to include patient carers, who lack professional medical knowledge but are also solely responsible for patients' daily needs, and as such are key "healthcare stakeholders", defined as follows:

Healthcare stakeholders

- **General Practitioners:** They are responsible for handling acute problems that present themselves as abnormalities of chronic disease. They have a high workload and very little time with patients, relying on communication with nurses
- **Practice Nurses:** They run chronic disease management clinics, where their role is to help patients manage particular conditions. They have a single specialised role in these clinics, and have more time with patients but a similarly high workload
- **Carers:** They are responsible for the day-to-day care and wellbeing of patients, and may be parents, other relatives, or professionally employed carers.

Patients themselves for whom remote self-monitoring is potentially useful are also diverse in their needs and capabilities. Interviews established that patients managing chronic illness, the *participants* in a clinical implementation, generally have one or more of the following properties:

Table 7.8
Designs and data collection methods of psychology researchers

Research Design	Location	Data type	Data collection	Barriers
Descriptive	Field	Cross-sectional Longitudinal	Paper surveys	Compliance, intrusive
			Online surveys	Compliance, bias
			Interviews	Dishonesty, influence
			Diary studies	Compliance
Correlational	Field Lab	Cross-sectional Longitudinal	Online surveys	Compliance, bias
			Diary studies	Compliance
Experimental	Lab	Cross-sectional	Online surveys Custom software	Recruitment difficulty Ecological validity

Chronic disease patients

- **Elderly:** Chronic diseases such as Type 2 diabetes, asthma, hypertension and heart disease are most often contracted in old age. Elderly patients tend to have memory problems and aversion to technology
- **Comorbid:** Patients who are managing a chronic disease frequently have comorbidities. For example, diabetes is often related to high blood pressure. These patients may attend multiple management clinics and must remember to take different medications and attend appointments
- **Knowledgeable:** Chronic disease patients are generally highly knowledgeable about their diseases, and are proactive in self-monitoring. They may be in need of more regular assurance

In summary, researchers and their participants, as well as clinicians and their patients, are diverse in technical skill, their knowledge and their difficulties in working practice. While previous research has identified barriers that clinicians might face in employing EUD, this is the first attempt to understand who clinicians and patients are in relation to chronic disease management. Additionally, while previous systems have been deployed for EUD of experience sampling applications, they have offered less insight into who might use such applications and how.

7.4.2 Overcoming difficulties

Interview feedback related to perceived usefulness represents the barriers that Jeeves could overcome in both research and practice. Unique benefits were identified for both domains, but also commonalities, as shown in Table 7.9.

7.4.2.1 Researchers

Social psychology researchers use a variety of data collection methods, related to the three main types of research design, namely *descriptive*, *correlational* and *experimental*. Table 7.8 provides an overview of the collection methods used in each of the three designs, listed with their primary disadvantages as reported by the researchers. In particular, it can be seen that diary studies are used in descriptive and correlational research. Experimental research methods, while popular for their scientific rigour, suffer from a lack of ecological validity, as well as the difficulty of recruiting a suitable sample size to take part in a lab experiment. Diary studies, and by extension ESM studies, have little applicability in experimental research.

Correlational research takes place in both lab and external environments. All researchers had conducted research using cross-sectional online surveys, citing difficulties in getting participants to complete them, as well as the associated recall bias. A descriptive design had also been used by all five researchers. Of these, paper and online surveys were most popular, with the former being seen as intrusive, and the latter taking place too far from the event of interest. Diary studies had again been used to gather longitudinal information, but suffered from compliance issues.

Generally, diary studies can alleviate many barriers such as recall bias and ecological validity, but still suffer from compliance issues. Thus, Jeeves was seen by researchers as a way to improve compliance through remote monitoring, and the ability to trigger at appropriate moments.

7.4.2.2 Clinicians

In clinical practice, the barriers to effective chronic disease management are also related to the structure of assessing patients. In GP appointments, acute issues are given priority, leaving little time for the discussion of chronic disease management or general health promotion. While the symptoms of a presented issue are treated, repeated appointments may be necessary for recurring symptoms that come from poor management of the underlying cause.

In chronic disease management clinics, more time is spent helping patients manage these chronic conditions. However, follow-up appointments may occur as infrequently as once a year. Patients engaged in self-monitoring are often the “worried well” who seek constant reassurance from healthcare professionals. In contrast, patients who are not proactive in following up with their

Table 7.9
Summarised benefits of both clinicians and researchers

Clinical Practice	Research	Benefit
Patients receive instant feedback	N/A	Patient reassurance
Patients more independent	N/A	Patient empowerment
Patients empowered to self-monitor	Participants engage in self-monitoring	Intrinsic benefit
Clinicians track patients' data	Researchers can monitor compliance	Real-time information
Clinicians acquire more accurate data	Researchers acquire in-the-moment data	Information accuracy
Time saved through reduced appointments	Time saved through remote research	Work time saved
N/A	Researchers acquire contextual info	Information richness

clinician may, as expressed by one clinician, *"fall through the clinical care net"*. Without cues to take medication, they are likely to develop acute problems that require more urgent attention.

Clinicians recognise that Jeeves could help dependent patients manage their conditions, by sending reminder prompts for important management activities. If any form of self-monitoring does take place, this could be reviewed in clinics to determine factors that relate to relapses. For "worried well" patients, clinicians were enthusiastic about the prospect of instant automated feedback that would be returned in response to the patient submitting their readings. This would minimise the number of unnecessary reassurance appointments.

For patients otherwise capable of self-monitoring independently, clinicians suggested empowering these patients to track holistic aspects of their health such as stress, diet, and exercise. This could therefore reduce the number of appointments related to acute issues, which a Jeeves app could allow motivated patients to do effectively.

The advantages are summarised in Table 7.9. In particular, time saved, information accuracy, information immediacy, and participant benefits were the advantages perceived in both domains.

7.4.3 Necessary features

The barriers faced by stakeholders in their research and practice are indicative of the potential advantages that an ESM app deployment could have. However, these advantages are not unique to an EUD approach – a professionally developed, bespoke app could also potentially overcome these barriers. Thus, the necessary features of Jeeves for supporting its adoption are those that would not be available in a static, purpose-built ESM app. In Jeeves, end-user developers are afforded tailoring opportunities at two levels, namely:

1. The **app** level - being able to tailor an apps' function **between** different patient groups and research questions
2. The **individual** level - allowing app behaviour to vary **within** patient groups or research questions

Tailoring at the app level can be observed as an absolute necessity at the core of EUD, given the diversity of chronic conditions that could be monitored, as well as the variety of research questions that researchers wish to answer. Some diseases or research studies require regular readings and assessments to be taken, others require reminders for daily management tasks, and some may also require advice or feedback messages to be given under certain conditions.

Tailoring at the individual level is also considered an important requirement. Within each chronic condition, patients exhibit a variety of technological skills, clinician dependency, and knowledge of their conditions, such that there is never a "one-size-fits-all" protocol for any given disease. From a psychology researcher's perspective, the ability to tailor assessments to locations relevant to a participant, or to times where they are available to take a survey, is seen as a useful means of improving compliance, as well as capturing uniquely significant information.

While Jeeves has support for the majority of identified functionality, a requirement noted in both domains, but not currently implemented, was that of a collaboration framework where support would be in place to ensure that various stakeholders (such as research groups or clinical staff) could cooperate on a single protocol and share understanding.

7.4.4 Organisational factors

The organisational factors influencing adoption in both psychology research and clinical practice have a number of similarities, such that adoption of EUD platforms for experience sampling apps are contingent on the following *facilitating conditions*, illustrated in Figure 7.4 in relation to the stakeholders whom they influence:

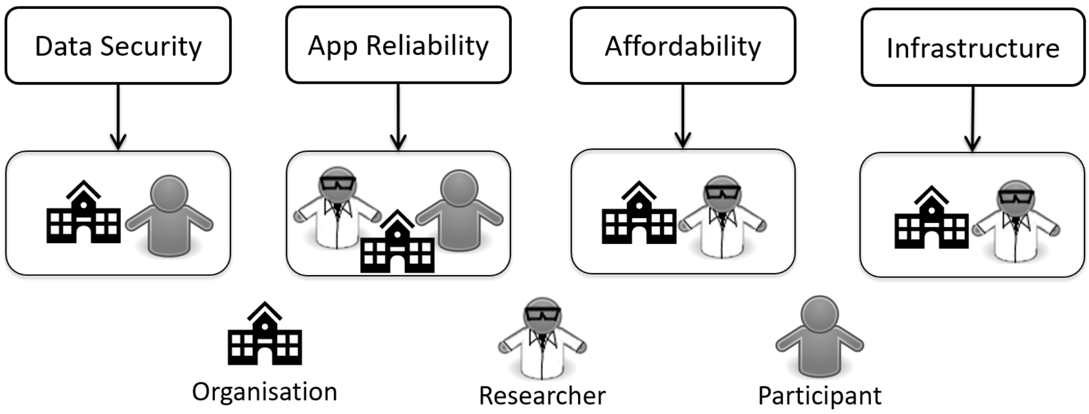


Figure 7.4
Organisational factors and how they affect organisation, researcher and participant

7.4.4.1 Data security

A prerequisite of ethical approval of any new technology involving participant information, data security was referenced by all clinicians and researchers in interviews. Stakeholders will require knowledge from the meta-designer of where data is being stored, the encryption process involved, and who will ultimately have access. App security will clearly have a significant impact on the **participant**, whose sensitive information could be leaked to malicious sources. The reputation of the **organisation** is also significantly affected by a security breach.

7.4.4.2 App reliability

Another prerequisite of ethical approval, and unique to *public-outward EUD* [142], it is of high importance that developed apps will function as expected in the field. To ensure this is the case, it is the responsibility of the meta-designer to thoroughly test the individual components from which an app can be built. However, it is also vital that the meta-designer prevents end-user developers from creating potentially harmful apps. Regarding Jeeves, syntax errors are eliminated through its blocks programming paradigm. However, *semantic* errors, such as infinite loops or inappropriately high-frequency sampling, are still possible. The reliability of an app impacts on the **organisation**, for the same reasons described above. It also influences the **developer**, who will not obtain their required data, and the **participant** who may stop using, or indeed be put at risk by an app that malfunctions.

7.4.4.3 Affordability

The cost of an EUD-ESM tool must be appropriate for the benefits that it provides. This is particularly salient at the individual **researcher** level for whom, reliant on external funding,

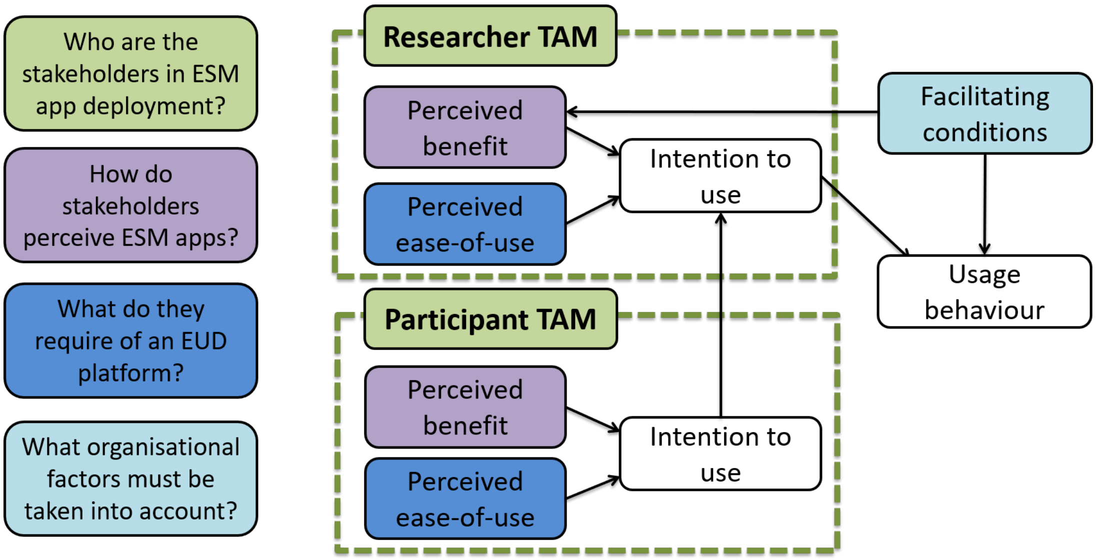


Figure 7.5
A layered Technology Acceptance Model, synthesised from interview feedback

affordability is likely to be a pivotal factor influencing adoption. In a healthcare setting, provided that the organisation approves the process and funds the application, affordability is less of an issue. However, in this case the scalability of the tool's storage mechanism must be considered. If deployed at a national level, patient numbers could reach hundreds of thousands, such that high capacity storage costs must be taken into account by the **organisation**.

7.4.4.4 Existing infrastructure

The *existing infrastructure* within an organisation refers to the processes currently in place and how an EUD-ESM tool will fit with these. While a transformation of working practice may eventually be possible with EUD, the adoption of an EUD tool will initially require it to be seamlessly *integrated* with working practice. Psychology researchers, for example, desired that Jeeves should work with existing software for participant recruitment and statistical analysis. Clinicians also desired that Jeeves should integrate with existing patient management software, and in particular, that they would be provided with a reduction in their standard workload to allow time to incorporate Jeeves into their everyday practice. Thus, modifications to existing infrastructure will impact on the **researcher** and the **organisation**.

7.4.5 Extending the Technology Acceptance Model

While researchers and clinicians have their model of technology acceptance, this is to some extent separate from the acceptance model of their participants and patients. Perceived benefits may overlap, but some are mutually exclusive. Further, perceived ease-of-use factors are separate, given that researchers and participants interact with two different interfaces. The models are linked, in that researchers will only consider adoption of EUD-ESM technology if their participants would be willing to use it. Thus, a layered model of technology acceptance was derived, and illustrated in Figure 7.5. In this figure, research questions pertinent to this chapter are colour-coded to the model constructs that they address.

- In addressing **who** the stakeholders of ESM apps are, only two distinct groups emerged in both domains, specifically researchers/participants, and clinicians/patients. As such, the model has two distinct layers, and could conceivably be applied to other examples of Public EUD with two communities of end-users.

- Both stakeholder groups have exclusive **perceived benefits** to the introduction of Jeeves that do not overlap. For example, clinicians would value the time saved through reduced appointments, whereas patients would feel intrinsic benefit from empowerment to manage their conditions. Further, researcher benefits of real-time study monitoring and contextual data acquisition will not be shared by study participants, who are likely to only perceive benefit through financial remuneration.

- Requirements of the EUD platform are related to the **perceived ease-of-use** construct. In approaches where the Software Shaping Workshop approach is employed, each stakeholder uses a different "workshop" application, with their own distinct usability issues. In short, the usability issues of Jeeves discussed in the previous chapter are unrelated to potential usability issues in the developed JeevesAndroid app

- The two layers are linked through the separate **intention-to-use** constructs, whereby a researcher's intention to adopt EUD technology is contingent on their participants' intention to use developed apps. Researchers were clear in their interviews that participants' ease-of-use and acceptance of JeevesAndroid were primary factors to their own adoption of Jeeves.

- Finally, irrespective of the intention to use of both researcher and participant, factors pertaining to a researcher's organisation (termed **facilitating conditions** in this extended TAM) may ultimately decide whether the technology is adopted. In interviews with clinicians, many stated that the NHS simply would not allow certain apps to be incorporated into practice, which clinicians had no power over. These factors also feed into researchers' **perceived benefit**. Interviewed clinicians suggested that, even if apps created using Jeeves were permitted in practice, if the NHS did not allow time for clinicians to dedicate to remote monitoring of patients, this would ultimately be less beneficial than the current practice.

7.5 Chapter Summary

Interview feedback from clinicians and psychology researchers gave necessary insight into the factors that would influence the adoption of Jeeves, or indeed any EUD-ESM tool, in their respective domains. The features implemented in Jeeves and evaluated in the previous chapter were found to be useful for applications in research and practice, and uncovered additional requirements necessary for adoption by a wider range of clinicians and researchers.

As well as clarifying the properties of an ideal EUD-ESM tool, interviews enabled the understanding of work practices within two distinct domains. Interestingly, despite the clear distinctions, from the perspective of Jeeves there were also clear overlaps in perceived usefulness, ease-of-use requirements, and organisational factors related to the acceptance of such technology into work practices.

CHAPTER EIGHT

OBSERVATIONAL RESEARCH

Ideas in this chapter have been published in the following paper:

D. Rough and A. Quigley, "Towards end-user development for chronic disease management," in *Designing Technologies to Support Human Problem Solving*. 2018.

Usability studies and interviews conducted in the previous two chapters gave insight into both the ease-of-use and real-world usefulness of Jeeves. Chapter 6 demonstrated that the complex features of Jeeves can be quickly understood by those with no prior programming experience. Further, Chapter 7 showed how both researchers and clinicians would value the use of an EUD-ESM tool such as Jeeves. However, in order to determine how it might realistically be used, and address usability issues in the field, it was necessary to observe the intended users of Jeeves in their working environments. This chapter describes the research conducted in order to do so.

First, two case studies are described where Jeeves was employed by psychology researchers at a local university, in order to conduct an ESM study for answering their own research questions. The two researchers used Jeeves in different ways, and thus perceived their own benefits and drawbacks with regards to its utility and usability. Results from the second case study are limited, however, as these researchers are currently at an early stage of piloting their developed app. However, the first study serves as validation that Jeeves can be employed effectively in practice. Furthermore, both highlight the potential for Jeeves to be applied in different contexts, partially addressing the final research sub-question of *"How do researchers and clinicians employ an EUD-ESM tool in practice?"*

Following this, analysis of an observation session conducted at a local clinic is discussed, where two chronic disease specialist nurses and a GP were observed as they conducted routine practices. Although Jeeves was not introduced during this session, the ethnographic approach provided further insights into possible deployment in a real-world situation. The results, while unable to disclose practical experiences, address the research sub-question posed in the previous chapter - *"What barriers do researchers and clinicians face that an EUD-ESM tool could alleviate?"*

8.1 Research Methods

The research methods used in this chapter enabled analyses of both Jeeves in its intended context of use, and of this usage context itself, with insights complementing those of lab-based usability studies and interviews. Foremost, case study research provides the level of detail required to answer research questions involving individuals interacting with a technology in their everyday environments. Robert Yin describes a case study as follows:

> Case study definition
>
> "A case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident. In other words, you would use the case study method because you deliberately wanted to cover contextual conditions - believing that they might be highly pertinent to your phenomenon of study" [206, p. 13]

In terms of end-user development, the phenomenon of interest is the EUD tool, and the context is the working practices that its introduction is designed to facilitate. Usability issues could also be investigated with greater ecological validity, given that the necessary tasks are grounded in researchers' own questions. Three categories of qualitative case study goals are further defined by Yin as follows:

- **Exploration**: understanding novel problems or situations, often with the hopes of informing new designs
- **Explanation**: developing models that can be used to understand a context of technology use
- **Description**: documenting a system, a context of technology use, or the process that led to a proposed design

Case studies in computer science research are applicable to further categories. In software engineering, Runeson and Host define **Improving** case studies, as those *"trying to improve a certain aspect of the studied phenomenon"* [207, p. 135]. In HCI literature, Lazar et

Table 8.1
Data sources used for each research question

Question	Case study sources	Triangulation source
Perceived ease-of-use	**Direct Observation** Created specification Email dialogue	Usability studies (Chapter 6)
Perceived usefulness	**Semi-structured interview** Email dialogue	Interview studies (Chapter 7)

al. further define **Demonstration** case studies as those *"showing how a new tool was successfully used"* [208, p. 160]. Interviews conducted in the previous chapter represent the exploration phase of this research, with which a greater understanding was obtained of the problems facing psychology researchers in conducting longitudinal studies, and those facing clinicians in employing patient-centred technology. Explanation-focused research was also performed to attempt to define an extended Technology Acceptance Model for EUD.

By these definitions, the case studies documented in this chapter aim to *describe*, *demonstrate* and *improve* the use of Jeeves in its natural context of psychology research. Both cases describe how Jeeves was introduced to researchers and the process by which the studies were planned, developed and eventually deployed with Jeeves, with feedback utilised for its consequent improvement.

Case studies can also be described as **intrinsic** or **instrumental**. Intrinsic case studies are those in which the results are only applicable to the individual study context, while instrumental case studies have results that can be more broadly applied to many potential users. The two case studies conducted are instrumental, in that they generated insights that could be applied to researchers in more general terms. They cannot, however, be applied to clinicians, whose work practices and potential use cases of Jeeves are significantly different. Nevertheless, even between psychology researchers, there is considerable diversity in research goals and methods. As such, these studies are an example of **theoretical replication**, where cases demonstrate contrasting results for predictable reasons [206]. While both cases involved researchers collaborating to create a study specification for their research question, there are notable differences: studies took place over different time periods, created specifications contrasted in complexity, and the process of designing and testing the study differed between researchers. As such, both case studies provided a unique example of Jeeves in practice.

Case studies typically consist of multiple data sources. In both studies, data was obtained through interviews, observations, email correspondence, as well as researchers' developed app specifications. Sources used in answering each research question are summarised in Table 8.1.

In addition to case study research, **naturalistic observation** was employed at a local clinic. With roots in social psychology for understanding contextual influences of behaviour, observation can be used in HCI to understand how software is used "in the wild", or current problems that introduction of software could potentially solve, as in this case. The use of ethnography in the design of new systems is articulated well by Blomberg et al. as follows:

"While the ethnographer is interested in understanding human behavior as it is reflected in the lifeways of diverse communities of people, the designer is interested in designing artifacts that will support the activities of these communities" [209, p. 123]

In summary, both research methods, in their proximity to the context of use, exposed factors pertaining to adoption of Jeeves as a real software artifact in a community of practice, rather than just a conceptual idea. The practical adoption of a new technology takes a staged approach, described in the following section.

8.1.1 Diffusion of Innovations Theory

Rogers' "Diffusion of Innovations Theory" proposes that there are five stages in the adoption of a new innovation, in this case a new technology [210]. These stages are summarised by Rogers as follows:

1. **Knowledge:** Person becomes aware of an innovation and has some idea of how it functions
2. **Persuasion:** Person forms a favourable or unfavourable attitude toward the innovation
3. **Decision:** Person engages in activities that lead to a choice to adopt or reject the innovation
4. **Implementation:** Person puts an innovation into use
5. **Confirmation:** Person evaluates the results of an innovation-decision already made

In the previous chapter, interviews conducted with psychology researchers and clinicians represented the **knowledge** phase of Jeeves. In demonstrating Jeeves and directly asking for qualitative feedback on its utility, the process of **persuasion** was also initiated, where researchers developed their attitude towards the system. Indeed, at this stage, two of the researchers were sufficiently persuaded to make the **decision** to use Jeeves in practice. This is consistent with the *trialability* factor of innovations, whereby potential users will be more enthusiastic to adopt an innovation if a trial period is permitted before commitment to adoption. The innovation-decision process is thus sped up by minimising the initial cost to researchers of doing so. Blackwell's Attention Investment Model theorises that the decision of whether to adopt a technology involves an end-user weighing up the costs and benefits of the adoption [211], thus the benefits are significantly more prevalent if the commitment cost can be kept low.

Following the **implementation** process and use of the innovation, the end-user can then decide whether or not to continue its adoption over a sustained period of time - the **confirmation** stage. However, evaluating whether the adoption of Jeeves would be *sustained* over a period of time is arguably the incorrect evaluation to be making. Tetteroo and Markopoulos directly relate Rogers' theory to the evaluation of EUD tools, and suggest that determining the active sustained use of an EUD technology might not be appropriate if EUD practices are going to be infrequent [212]. Given that the duration of an experience sampling study is relatively short, particularly in comparison to phases of ethical approval, recruitment, and data analysis, it is perhaps unlikely that Jeeves would be adopted as part of *everyday* working practices. Therefore, the focus of this chapter is predominantly on the **implementation** phase of Rogers' theory. As described by Tetteroo and Markopoulos, the case studies of this chapter were designed to:

"evaluate the EUD practices that end users develop, the role that EUD starts playing in the context in which it is deployed, and most importantly, the extent to which the EUD practices help the end user to achieve his or her goals" [212, p. 226]

However, the Diffusion of Innovations Theory did not directly apply to the clinical observations - Jeeves was not introduced to the practice nurses and thus the **knowledge** stage was not initiated. However, Rogers' model is still of relevance in this case. In simply observing without applying an innovation, this provides a greater understanding of the external, organisational factors relevant to its adoption. In psychology, the innovator is effectively the individual. The use of ESM in psychology research is prominent, and the individual researcher is effectively at liberty, within the constraints of his or her research budget, to decide whether to adopt a technology. However, in clinical practice, the innovator must be the *organisation* - in this case, the National Health Service, which presides over the technology solutions that are adopted at both a local and nationwide scale. With a small number of exceptions [32, 33, 79], there appears to be very little empirical evidence of the benefits of ESM in clinical practice. The interviews conducted in the previous chapter certainly provide clear suggestions from individual clinicians on the factors that would influence the adoption of Jeeves. However, these individuals cannot speak for the organisation as a whole, and thus the *understanding* process is more complex and requires greater investigation in its potential context of use.

8.2 Case Studies

The case studies were conducted with two psychology researcher interviewees in Chapter 7. Their intrinsic desire to use Jeeves in their own research made them ideal representative cases. Moreover, conducting in-depth studies with previous interviewees allowed triangulation of the case study data with that of their interview feedback to answer a primary research sub-question of this thesis, namely: *"How can researchers and clinicians employ an experience sampling app end-user development tool in practice?"*

This question was partially addressed in Chapter 7, where the preliminary feedback of researchers was gathered through semi-structured interviews, suggesting how they would like to use Jeeves in practice. To a lesser extent, it was also addressed in Chapter 6, where usability issues of Jeeves encountered by non-programmers while completing tasks representative of psychology research studies were obtained. However, both usability studies and interviews lacked the ecological validity of real-world usage, and could not identify the additional emerging requirements in doing so.

In the usability studies described in Chapter 6, assumptions were made about the typical study specifications that psychology researchers would create. While these assumptions were informed by publications in psychology journals that utilised ESM, the constraints of lab usability studies meant that too many assumptions were made about researchers' practical application of Jeeves.

The interviews documented in Chapter 7 also have limitations with respect to real-world utility of Jeeves. While knowledge of researchers' use of software in practice was obtained through these interviews, this information was anecdotal. Software is used sporadically in practice when it is required, and is not typically used on a daily basis. Moreover, researchers seldom introduce new software into their research practice, typically training themselves to use university-licensed software, such that capturing experiences on their learning curve and experience with using new software was difficult.

8.2.1 Research questions

Thus, investigating how researchers used Jeeves in practice to perform tasks directed towards their own goals exposes real-world usability issues and how they were overcome, as well as their perceived usefulness of Jeeves during the design and deployment process. Although no refutable hypothesis was defined prior to conducting the case studies, it was expected that the researchers would successfully use Jeeves to conduct their studies, but would require continuous support in doing so. Moreover, it was expected that Jeeves would not fully address the researchers' requirements from the beginning of the study, and instead a meta-design process would evolve whereby researchers would request new features in response to their particular needs. In summary, the case studies were used to determine the following:

- The **usefulness** of Jeeves in fulfilling researchers' goals - how effective Jeeves was in letting them run an experience sampling study, what challenges they faced, and what factors would influence them to decide to use Jeeves, and further, to continue to use Jeeves.

- The general **usability** and intuitiveness of Jeeves - how quickly researchers learned to use the functions of Jeeves to achieve their goals, what problems occurred and how they resolve them, and how its usability compares to other software they commonly use.

- The perception of a **meta-design** model and their roles in a culture of participation. What researchers would desire in collaborating with each other, whether and how they would use peer support, and how open and modifiable Jeeves should be to the specific needs of individual researchers. This question transcends the usefulness and usability of Jeeves, considering meta-design as an overall framework that would ultimately give end-user researchers the power to influence the usefulness and usability themselves. This was an important consideration given that it was not previously possible to evaluate this in the usability or interview studies, and could only be evaluated in practice.

As an active participant in email discussion and collaboration, I will further refer to myself in the first-person where necessary for clarity during this section.

8.2.2 Data collection methods

Data collection methods were identical in both of the reported case studies. The primary means of communication with the researchers was through email correspondence. Troubleshooting application issues, discussing study design and general usability feedback on the Jeeves application was mostly communicated through a series of email threads.

In the role of a meta-designer, I received feedback from the researchers through email correspondence, Skype discussions, and face-to-face meetings, implementing requested extensions as required, which the researchers then integrated into their study specification. Performing this role, as opposed to passively observing researchers developing their study with the functionality available, allowed both myself and the researchers to experience how a meta-design framework would function in practice, whether it was a necessary factor for the usage of Jeeves, and how the meta-design process could be improved. Thus, engaging in this role, rather than simple passive observation of researchers, maintained a degree of ecological validity.

Indeed, using the terminology introduced by Cabitza et al., my role additionally encompassed that of a "maieuta-designer" [142]. In considering *meta-designer* as a technical role that primarily involves the design and implementation of the EUD tool, the *maieuta-designer* is responsible for developing the social conditions for domain experts to become end-user developers. As such, they support and encourage EUD activities, thereby growing the self-confidence of these domain experts to increase their independence from professional developers. Through liaising with the case study participants, assisting with technical issues, suggesting workarounds, and facilitating pilot studies, my role transcended that of simply the developer of Jeeves. Although I would still remain active as a meta-designer in future studies that researchers wished to run, my role as maieuta-designer would instead gradually reduce as the self-efficacy of these researchers grew to enable them to engage in EUD activities independently.

8.2.2.1 Secondary sources

The secondary sources in this study were, primarily, the chain of email correspondence between the two researcher groups. Additionally, their developed Jeeves specification was inspected and used as a secondary source, and discussed with the researchers in subsequent interviews at the end of each case study.

8.2.2.2 Direct observation

A single instance of direct observation took place near the beginning of each case study, where researchers were observed working together to design their study specification. During this session, passive observation, as opposed to active participation, was employed, in order to avoid influencing researchers' decisions in their design process. When questions were raised, researchers were directed to consult the tutorial video. However, uncertainties were also verbally confirmed that were not specifically addressed in the video.

8.2.2.3 Interview

Following the completion of the first case study, a semi-structured interview was conducted with these researchers, to elicit further feedback on their experience with Jeeves, the benefits and barriers of its use, and factors that would influence its future use in their research. In addition, my own role in the design and development process was queried, in order to understand what role researchers felt software developers should have in the design process. (This was not conducted in the second case study, which is ongoing at the time of writing.)

8.2.2.4 Observation

Observing the progress of each study was primarily achieved through email correspondence, whereby researchers would initiate contact to provide updates on the progress of their study, or to ask technical questions. Alternatively, researchers who had not done so for over a week were prompted to ask how their study was progressing. My role in the observation process was different in the two studies, however. In the first reported study, I was a "participant observer" as defined by Yin [206]. After the initial observation session, for the duration of the case study, I took an active role in assisting the researchers in resolution of design and implementation issues, adding new features as required, and facilitating the pilot study that was conducted. Thus, a dual role of *meta-designer* and *research collaborator* was performed, which created both beneficial conditions, and drawbacks as will be discussed.

In contrast, a more passive observation role has thus far been undertaken in the second study. Clearly, in active continuation of my role as a meta-designer, it is impossible, and indeed undesirable, to be divorced completely from the study. However, in this case, the secondary role is not as a collaborator, but instead as a source of technical support. This has advantages in alleviating issues, caused by the influence of direct collaboration in the first case study. However, this also has disadvantages in that it is more difficult to obtain details of the research process. This difference in the two case studies has resulted in what Yin terms as a "theoretical replication", where a further case study is conducted that *"produces contrasting results but for predictable reasons"* [206, p. 46].

8.3 Case Study 1 - ESM During Sport Events

The first case study was conducted in collaboration with a psychology researcher at a local university, and a collaborating researcher at a university in Germany. This case study reports on the experience of collaboration with these researchers, the design of a suitable study specification for the researchers' questions, and issues regarding testing, deployment and analysis of data collected with Jeeves.

8.3.1 Background

The researcher identified as P4 in Chapter 7 interviews, hereby referred to as Paul, and the collaborating researcher, who will be referred to as Oliver, have previously conducted a study regarding the shared experiences of fans at different types of crowd event (for example, football matches and music concerts). Paul expressed in his interview that ideally, he would like to collect data on the fluctuating feelings of identity in fans at these events, but explained the difficulties in doing so. In particular:

1. The practicalities of handing out and collecting paper surveys at such events was a physical burden on both supporters and the researcher.
2. Supporters were happy to share their experiences, but were suspicious of filling in consent forms, and were not interested in lengthy paper surveys: *"I found if you're going to a protest event or a football match or a music festival, people are there because it's their free time...they don't want to do paperwork"*.
3. Cross-sectional paper surveys did not take into account dynamic variations: *"We're interested in social context, changes, how people, cognitively appraise the situation...and if you just take a snapshot you're not entirely sure what's caused that or what the consequences are"*.

Both Paul and Oliver had an interest in capturing these experiences of fans at regular intervals during a sporting event. Their current knowledge of programming would not allow them to do this easily or cheaply, thus Paul saw how Jeeves could be used for this purpose and initiated the **implementation** phase of Rogers' Diffusion of Innovations Theory [210]. What follows is a documentation of this implementation, and an interview that evaluates the researchers' **confirmation** phase with respect to this theory.

8.3.2 Timeline

The following section documents the events that took place in the duration of this case study. A diary of meetings and communication was kept, augmented by the ongoing email thread. In particular, *technical* and *collaboration* issues were common throughout this process.

In November, following the interview with Paul conducted as part of the research in Chapter 7, email correspondence began with Paul and Oliver. Requirements of a potential study were ascertained, including suggestions for a possible pilot study, and that of a full-scale study. A Skype call clarified further options and time frames. It was decided that the goal was to conduct an experience sampling study with supporting fans of a popular sports team at Oliver's university during a live game, prior to which a pilot study would be conducted with local students watching a live football game on television. At this stage, the necessary documentation for ethical approval was organised, including the questions that were to be asked to participants, as well as advertisement material for the study in Germany. Documentation was provided detailing the security protocols in Jeeves, as well as who had access to the data stored.

In December, the original ethical approval documentation was submitted. Prior to this, various document drafts were circulated during the planning process. Brief notes were written up by

Paul and shared as a group document via email. This included discussion on the technical capabilities of the app, as well as proposed survey questions and when they should be sent. Updated survey questions were later sent in a separate document with annotations. Approval was eventually received on 20th December, prior to the university closing for the Christmas holidays. In early January, a further Skype meeting was held and recorded, during which the researchers watched a 10 minute video tutorial of Jeeves, and used the screen-sharing function of Skype to collaboratively design a study specification. Prior to the meeting, Paul circulated the ethical approval documents containing the questionnaire, participant information sheet and debriefing form, for reference during app design. Both the audio and video of this Skype call were recorded for later analysis.

8.3.2.1 Preparation for pilot study

During the call, collaborative completion of the study between Paul and Oliver was observed to be difficult through the screen sharing function of Skype, whereby Oliver would dictate survey questions from the plan document while Paul created the survey in Jeeves. Complications further arose when informed consent had to be implemented into the first survey. Initially, Paul copied the text from the consent survey, but it was then incorrectly formatted. It was suggested that providing participants with a URL link to an informed consent document would be simpler, which was agreed upon. An issue of uncertainty arose wherein a question required two numbers to be input, and had to be split into two separate (but closely related) questions.

> Previewing Issues
>
> **Paul**: Actually, shall we make this...
> **Oliver**: On one page?
> **Paul**: I'm not sure how that's gonna work with pages, if it's a new page per question, or what's gonna happen..
> **Oliver**: That would be good if it's on one page, right?

Here, the lack of preview functionality for surveys resulted in difficulty. Further, inability to duplicate similar questions also became an issue. Both Paul and Oliver made comparisons with *Qualtrics* - software they were both familiar with in their research:

Qualtrics Comparisons

Paul: Is there a way of previewing questions? I mean I guess it's kinda here, that'd be really useful. That's something *Qualtrics* does and it'd be quite useful. There's not a way of copying a question is there?
Oliver: Paul, you also use *Qualtrics* right? I think it has very...smart features, especially what you said, copying questions, preview of questions, and also these randomisation things, orders, stuff like that.
Paul: Yeah there's a lot of good stuff in *Qualtrics*, it can't do everything we want it to do, but in terms of user features it might be worth...not ripping it off but...
Oliver: Yeah I have to agree, but they sell this for a lot of money so...

Oliver then left, agreeing to install Jeeves and implement a section of the study. Paul suggested that a form of annotation would be desirable for communicating ideas to Oliver:

"...to add a comment, annotation...a note to yourself to say 'I've still got to do this' or 'remember to change that' or in a collaborative project, 'I'm not sure how this works' or 'what do you think of this?'...just to say 'Oliver this is for the half-time survey, just starting it for you, you finish it' "

Technical issues - Integrating informed consent and debriefing into surveys was identified as absolutely necessary prior to running the study with participants. At this stage, the lack of a previewing feature was also noted by both researchers.

Implementation of new features caused synchronisation problems, however. On two instances, a new feature was implemented, and the updated version of Jeeves uploaded. When a researcher with the updated version added the novel feature to a study, this made it impossible for the other researcher to access this study without first downloading the new Jeeves version.

Collaboration issues - It was observed that, at least for remote collaboration, communication would be difficult. Screen-sharing with Skype was cumbersome and forced researchers to switch between Jeeves and previously created design materials.

Following the Skype call, communication continued as before, through the shared email thread. The types of communication were varied, and interleaved within emails, including:

- Declaration of study updates (*I've now coded the pre-match and post-match surveys in the Jeeves app*)
- Requests for new features *(I totally agree with Paul that a way of copying questions would be really helpful*)
- Requests for bug fixes (*When I login the two projects I can see are 'test' and 'test2' but I am*

not able to open either of them. Any ideas how to proceed?!)
- Requests for collaboration activities (to myself and researchers) (*@Oliver - Can you now try and do the half-time survey?*)
- Clarification requests (*Are you able to see the data after entering dummy data?*)

Responding to emails in a way that separated roles as a meta-designer and a collaborator became more difficult. The different communication types were interleaved such that email responses to researchers contained elements of both roles.

8.3.2.2 Conducting the pilot study

On January 13th, the pilot study was conducted. Nine participants were initially recruited, during which they watched a live football game on television in the university's School of Computer Science. They answered survey questions prior to the match starting, at half-time, and at full-time, as was to be the case during the full-scale study. Participants were welcomed and a link was provided to the beta version of JeevesAndroid on the Google Play store. After downloading JeevesAndroid, participants were provided with instructions to access the study and complete the introductory survey. 10 minutes before kick-off, at half-time and at full-time, most participants received survey notifications on their devices. After completing the final study, they were thanked and received £10 compensation.

Technical issues - One participant had an incompatible device, running a lower version of Android than was necessary for the study. Another participant had privacy settings enabled on his device, so that he did not receive surveys at the same time as other participants.

One participant turned up particularly late, and by the time he had installed the JeevesAndroid app and initiated the study, he had missed the first trigger. The time for this trigger was adjusted through Jeeves so that it would be sent to his device, which meant that it was also sent to all other devices, causing confusion.

Participants were further confused about the various app permissions required. Previously, to install the JeevesAndroid app, all permissions that might be required in a study were requested on installation. It was necessary to explain to participants that this data was not being collected for the particular study, after which they agreed to continue the installation.

Collaboration issues - To ensure the validity of assessing use of Jeeves in practice, determining how the researchers would have resolved the pilot study issues in practice, or whether this would have ultimately caused abandonment of Jeeves, would have been useful.

8.3.2.3 Preparation for full-scale study

A Skype call was set up to discuss the results of the pilot study and to plan for the full-scale study in Germany. Only the audio of this call was recorded, as no use of Jeeves took place. In his testing of the app, Oliver noted the same issue that participants had experienced regarding the necessary app permissions, suggesting that this would need to be corrected. Given that it may not always be possible to meet participants to explain permission requirements, participants' initial experience is particularly important:

"Users usually don't like that, especially when it's something new...we cannot explain anything to them. We have to rely on the usability of the website and the app." (Oliver)

Oliver was responsible for recreating the pilot study, with survey questions written in German, and trigger times adjusted to key phases of the basketball game. Given that the design was otherwise identical, Oliver commented that a feature to simply duplicate the pilot study specification would have been useful. What emerged at this stage was the effort required to organise a full-scale study. The process involved arranging with the sports team owners, publicising the study on social media, sending out further emails to interested participants, and arranging reimbursement for participants' time (in this case, participants were entered in a prize draw for a gift voucher). Given the various organisation activities involved, as well as the researchers' other commitments, development activities were put on hold. After the last update by Oliver on January 23rd, no further updates were made until January 27th - one day before the full study - when a bug was discovered in which participants who had registered were already being sent study surveys.

Technical issues - The main issue following this Skype call was the need to translate the app into German. While the survey questions could be translated by Oliver (a native German speaker) there were many screens on the app that were hard-coded in English, requiring a tedious process to ensure that all hard-coded text fragments were updated. Additionally, a new version of the app had to be created.

Collaboration issues - While previous collaboration issues had been caused by the difficulty for researchers to communicate through Jeeves, the bug occurring one day prior to the study was an example of Jeeves failing to communicate to the researchers. The bug was only discovered by chance. Without any form of notification, researchers were unaware of any issues.

8.3.2.4 Running the full study

The full-scale study was run as planned; 40 participants initially signed up, and 30 completed every survey. Due to the variation of basketball match times due to fouls and timeouts, Oliver was present at the match to adjust the half-time and full-time surveys as necessary, to ensure

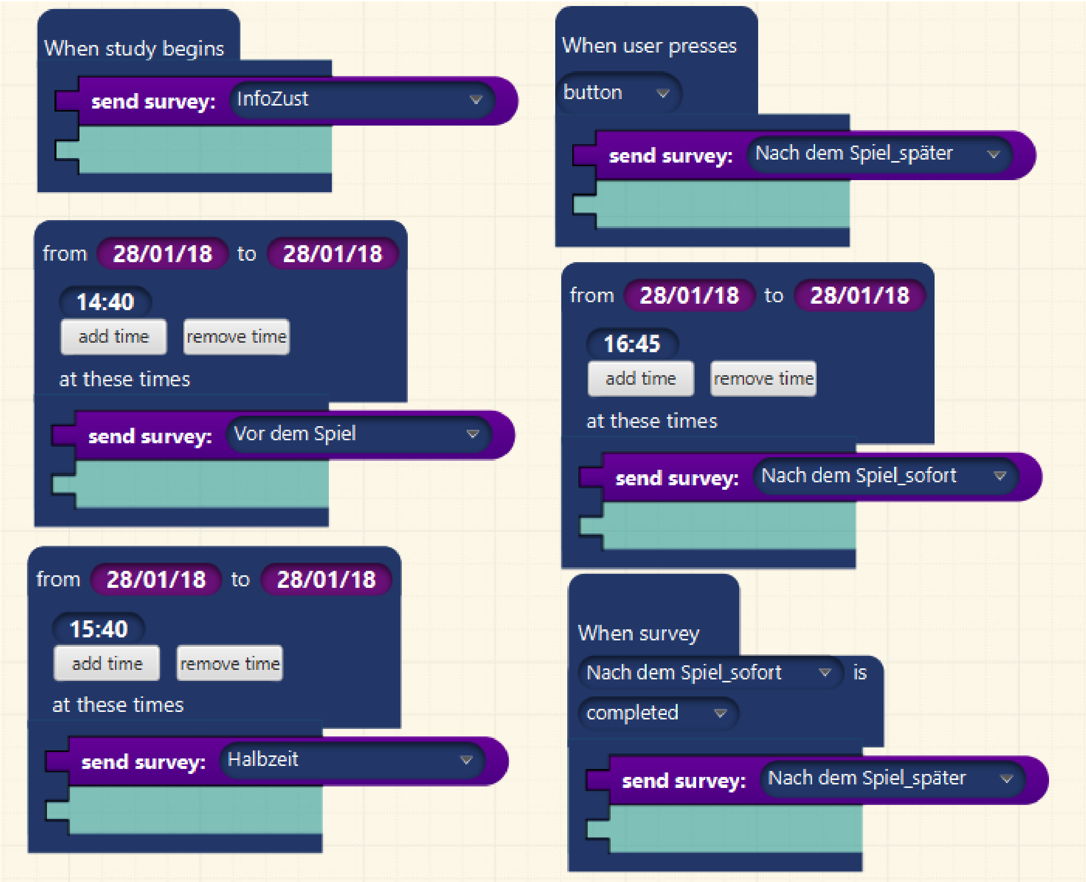


Figure 8.1
The final specification for the study conducted at the basketball game

that participants would receive surveys at the appropriate times. However, in the final study specification, Oliver had left the "Button Trigger" and button he created for testing purposes in the version of the app that participants downloaded. Some participants found this button and ended up completing the post-match survey too early. (This trigger can be seen in the top-right of Figure 8.1, showing that when the button is pressed, a post-match survey is sent.)

The final post-match survey was instrumented with Likert scale questions, ranging from 1-7, regarding the ease-of-use of different app functions, and a final free-text question to discuss any positive or negative aspects of app usage. The questions and their average ratings (out of 7) are as follows:

Please use the sliding scales to describe the ease-of-use for the following.

- Signing up for a study $(5.97, SD = 1.08)$
- Responding to a prompted survey $(5.71, SD = 1.35)$
- Manually starting a survey $(5.32, SD = 1.68)$

- Free text questions $(4.97, SD = 1.78)$
- Multiple choice questions $(5.97, SD = 1.11)$
- Sliding scale questions $(6.00, SD = 1.18)$

Although nothing significant emerged from these results, it was noted that the free-text questions were rated lower than other app functions, with a higher deviation showing some participants expressed a strong dislike of such questions. One explanation is that these questions require more time and effort to fill out, thus distracting participants from the atmosphere. This was reflected by participants' one-word answers, or complete omission of open-ended questions altogether. Other negative feedback was related to the presence of the Jeeves icon in the navigation bar, which participants disliked: *"ALWAYS runs in the background. That's annoying"*

Further, participants were unsure of how to proceed at the welcome screen, prompting them to explore the interface and accidentally trigger the post-match survey: *"The design is not particularly appealing, additionally at the beginning I didn't know where to find those surveys"*; *"Confusion at the beginning about what to start when"*; *"The design is not appealing and the starting page is not user-friendly: not really self-explanatory"*.

8.3.3 Interview

Following the full-scale study, a semi-structured interview was conducted with the researchers to obtain their overall impression of Jeeves, usability issues that would be factored into their future use of Jeeves, the utility of Jeeves in achieving their research goals, and the practical aspects of meta-design. For example, the interview aimed to discern their overall experience of collaboratively running the study, their view of a meta-designer's role as a collaborator, and how roles should be separated in future. As with previous communication, Paul was in the room and Oliver was present through Skype. The audio of this interview was again recorded for later analysis.

The interview protocol is included in Appendix C, Section C.4. Specific points raised in the interview, which support or contradict previously derived adoption factors, are described in the Discussion section below. Although affordability and context-aware triggers were also highlighted as useful in the interview, this was mostly a reaffirmation of the points Paul made in his first interview. Moreover, as they were not observed in practice in this case study, they are not included in this discussion.

One unique point made in the interview that was not previously discussed in Paul's interview, was that of social influence in the decision to adopt a new technology. Paul explained that published articles were a useful means of promoting a new technology:

Table 8.2
Observations made in Paul and Oliver's case study that relate to design factors in Chapter 6

Observation	Supporting Design Factor (Chapter 6)
Researchers wanted to annotate their created blocks for clarity	**Use text to distinguish graphics**
The blocks were manually ordered by researchers to show chronological order	**Have 'tidying up' features**
The full study specification had to be replicated from scratch	**Allow reuse of examples**
Paul was not able to preview his created app design without an Android phone	**Bridge the gulf of evaluation**

"For me I guess whenever there's a new bit of software, so I'd have to find it first of all, so it has to be visible through a web search or something...and a good way of doing that of course is for us to publish stuff...that other social scientists will find in their web searches"

Overall, Paul and Oliver were happy with the study as a proof-of-concept test of the app's functionality. While the data was interesting, the sample size of 30 and the issues related to surveys being sent too early or triggered accidentally by the users, meant that it was not rigorous enough for full publication.

8.3.4 Discussion

The case study exposed how psychology research studies are conducted, and factors that would influence EUD introduced in this domain. Previous EUD-ESM tools do not appear to have taken into account the informed consent and debriefing required, for example. Indeed, very little research has been done into introducing ESM into practice like this, an exception being the *Tempest* platform by Batalas and Markopoulos [213], and *PartS* by Ludwig et al [101]. Neither of these report on case studies, however.

8.3.4.1 Usability

The usability of Jeeves was difficult to evaluate comprehensively - Paul and Oliver did not use the more complex features such as attributes and tailoring based on survey responses. However, both researchers reported that it took very little time with Jeeves to achieve the goals of their study. Thus, within their specific constraints, the time taken to learn Jeeves was insignificant in comparison to the time saved. However, one time-related issue occurred when Oliver was forced to re-implement the study specification for a German version of the app. The design factor to ***"allow reuse of examples"*** was pertinent here.

Quality-related issues included Paul's frustration at the inability to preview his created app, causing a lack of confidence in the app's reliability. As in previous usability studies, the fact that Jeeves does not ***"bridge the gulf of evaluation"*** was a salient issue. Ease-of-collaboration was a primary usability issue that did not arise in previous evaluation, thus a corresponding design factor was not generated for this in Chapter 6. However, the researchers made suggestions that fit with previously derived factors. For example, both desired the ability to automatically order triggers for a *"clean layout"*, as expressed in the design factor ***"Support 'tidying up' features"***. Further, researchers wanted to add their own annotations to provide additional explanation to the other researcher, consistent with ***"Use text to distinguish graphics"***. Beyond the use of *internal* text to define a block's function, researchers should be allowed to add their own *external* annotation text. A summary of these observations is given in Table 8.2. One limitation of the case study was that it was not possible to fully observe the researchers using Jeeves. Since no usage behaviour was logged, the only sources available were the initial direct observation session, their developed app, and their anecdotal feedback.

Despite the simplicity of their developed specification, Paul and Oliver were able to implement their design rapidly and with no need for intervention beyond clarifying the specific behaviour of certain blocks. The speed and perceived ease with which they were able to do so suggests that the blocks-based approach, while enabling more complex behaviours, is also not overwhelming for developers seeking a simple study specification. Further, despite the cumbersome screen-sharing Skype collaboration process, Paul and Oliver were able to discuss and alter the specification with no issues, by referencing blocks by name or type, as opposed to specific lines of code, for example.

8.3.4.2 Usefulness

The researchers created a simple specification that did not utilise much of the feature set considered important for Jeeves. While this was disappointing to an extent, it also fit with the requirement derived in Chapter 7 that ***"Jeeves must allow researchers to perform simple tasks"***. Furthermore, the review of studies in Section 2.2 highlights that the majority of studies have a similar structure to those of the researchers in this study. While researchers may eventually begin to utilise the novel features of Jeeves, for initial adoption, their research question will likely guide a simplistic study design.

With no collaboration functionality, it was occasionally the case that the researchers would modify

Table 8.3
Observations made in Paul and Oliver's case study that relate to utility factors in Chapter 7

Observation	Supporting Utility Factor (Chapter 7) Jeeves should...
Researchers' simplistic specification was guided by their research question	**...allow researchers to perform simple tasks**
Collaboration activities involved chains of emails and Skype screen sharing	**...support collaboration within and between research groups**
Participant attrition would have been caused by asking for too many permissions	**...be non-intrusive in participants' lives**
Paul had to spend time formatting collected data so that it could be used in SPSS	**...integrate with existing research technology**

the study simultaneously, causing the later save to overwrite the earlier save's changes. The requirement that ***"Jeeves should support collaboration within and between research groups"*** was supported here. When researchers are working in teams on different machines, collaboration support is of particular importance.

With regards to saving time, Paul appreciated that the data for all participants was integrated into one spreadsheet. However, because of the row-column format in which Jeeves stored this data, he had to spend time in first creating the necessary SPSS data file, explaining in an email that *"It took me an hour or so to build this manually, so if as much of this work could be done by the app as possible would be great"*. Additional unnecessary time would have had to be spent transposing Jeeves data into an SPSS-compatible format, supporting the utility factor that ***"Jeeves should integrate with existing research technology"***.

Finally, the pilot study was insightful in determining the quality and reliability of JeevesAndroid with regards to participants. Although all participants found the use of JeevesAndroid simple and straightforward, problems were caused by the large number of sensitive permissions asked for at the beginning of installation. A small minority of participants questioned this, and without my presence to resolve their concerns, it is likely that they would have not complied with installation. The requirement that ***"JeevesAndroid should be non-intrusive in participants' lives"*** refers not just to the burden of answering surveys, but also the feelings of intrusion caused by having to agree to location and sound recording. These observations are summarised in Table 8.3.

8.3.4.3 Meta-design

As a collaborator, I took responsibility for conducting the pilot study, and testing the researchers' developed application, responsibilities that would otherwise be taken by the researchers themselves. As such, this had a direct influence on the study's process, which would not have occurred in passive observation of the researchers independently managing the study. However, in doing so, greater detail was acquired of the research process, and the issues involved in collaborating, testing and running a study. Moreover, this collaborator role was necessary in order to ensure the study's full completion - the researchers were unavailable to run the pilot study themselves, in time to ensure that the full-scale study could take place.

Two disadvantages arose in that during Skype calls, engagement in the collaborator role meant that interesting facets of these calls were possibly missed, which would have been clearer in a passive observer role. As these Skype calls were conducted on Paul's laptop, they were not recorded. Further, and more importantly, it was unclear to what extent this had an influence on the process. Had the researchers been solely responsible for running the pilot study, testing the app, and clarifying their own confusion, their opinion of Jeeves may have been considerably lowered.

In a realistic deployment of Jeeves, the meta-designer role would not overlap with that of a collaborator. However, the meta-designer would still be of primary importance, in order to implement features as required in response to requests from a growing end-user community. A quote from Paul exemplifies how a lack of meta-design could result in quick abandonment:

"we were able to shape the design of this at very short notice so we were saying 'can you do this, can we change this'. I think otherwise we might have logged in the first time and thought 'well it doesn't actually do this', then try and move onto something else".

Although abandonment could be reduced at later stages of implementation by incorporating tutorials and FAQs, if primary functionality is not immediately available and cannot be requested, then researchers will default to more familiar software like *Qualtrics*.

8.4 Case Study 2 - ESM in the Menstrual Cycle

This case study describes the progress of a collaboration with a second psychology researcher at a local university, and her postgraduate research student. The focus of this case study was an experience sampling approach required for this student's final year research project. The researcher, who will be referred to as Deborah, participated in the interviews conducted in Chapter 7, and in doing so expressed enthusiasm for using Jeeves in a master's project. Deborah's area of research is in aggression, and she was particularly interested in understanding the contextual factors that influence feelings of aggression, and outward displays of aggressive behaviour. Thus, Jeeves was considered ideal for exposing contextual factors, outside the constraints of the traditional laboratory experiment.

8.4.1 Background

Unlike Paul, whose research interests in crowds mandated his use of field methods to collect data, Deborah's research methods were largely experimental. In researching aggression, studies would take place in an instrumented psychology lab, where participants would be exposed to a stimulus or a control condition, and their responses compared to determine a significant difference. There were notable issues with doing so:

1. The reliance on memory caused significant biases in self-reported data. Deborah explained an interesting phenomenon: *"if you ask people how often they've [struck or shouted at their partner] in the last month, and then you ask a separate group of people drawn from the same population how many times have you done this in the last year, the numbers will be basically identical"*. Retrospective recall bias thus reduces the reliability of data.
2. Ecological validity is also a barrier when conducting aggression research: *"There are experimental measures of aggression but they're not great. 'Cause there are very obvious ethical concerns with getting people to be aggressive to each other in the lab"*. As such, most of Deborah's data was forced to come from retrospective self-report.
3. Honesty in compliance with data collection is also an issue in this particular area of research. Irrespective of memory bias, participants may feel self-conscious about admitting acts of aggression towards their partner: *"people might not be, they might not want to be completely honest about their behaviour and so, anything that reduces the feeling of exposure from giving that kind of data would be a very useful thing"*

Deborah's master's student, hereby referred to as Lucy, is conducting a research project to investigate antecedents, consequences, and general variation of female aggression during the menstrual cycle.

8.4.2 Timeline

As previously explained, the length and detail of the timeline in this study were constrained by the researchers' own schedules, including obtaining ethical approval and recruiting study participants, as well as other delays including Christmas holidays and adverse weather circumstances. Additionally, as a passive observer and meta-designer, this precluded direct collaboration with

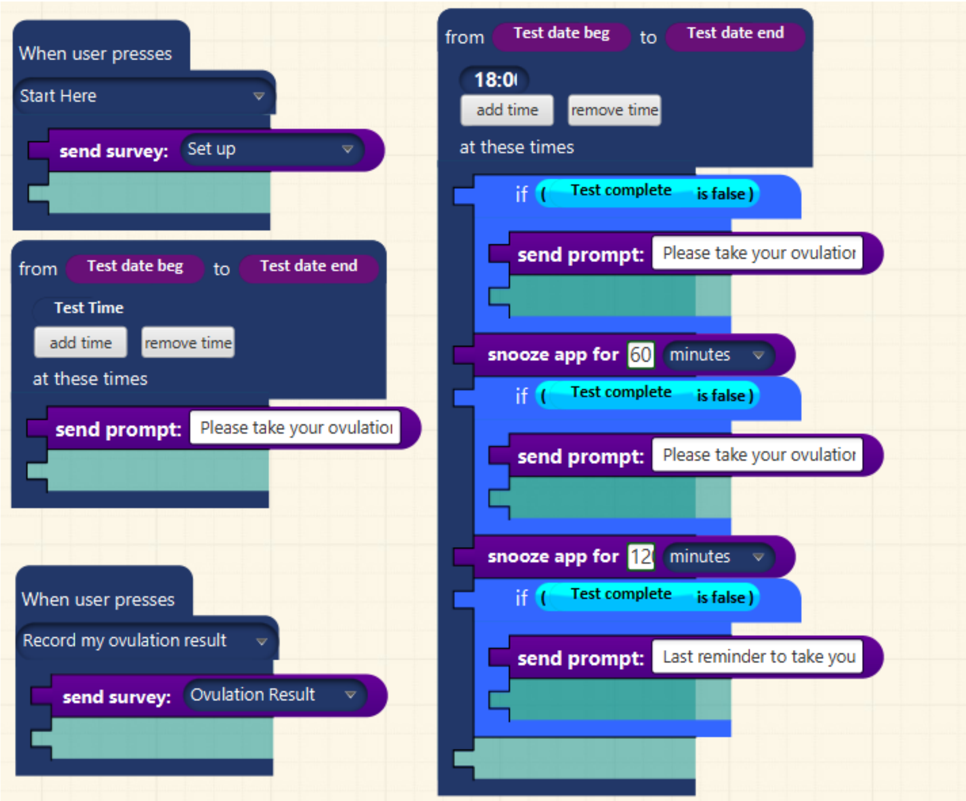


Figure 8.2
Deborah and Lucy's pilot study specification

Deborah and Lucy on this study, such that face-to-face meetings on the project were not observed, and as such, information related to study progress was sparse. However, the observation session wherein the researchers used Jeeves was particularly insightful as they utilised a wider range of Jeeves functionality than Paul and Oliver, whose study was relatively straightforward, and short in duration. Further, the researchers used some interesting workarounds to compensate for some of the functions that were not available in Jeeves, showing its flexibility towards adaptation.

8.4.2.1 Preparation for pilot study

In-person meetings were arranged in November to plan the preliminary tasks that would need to be undertaken prior to designing the study. As before, ethical approval had to be provided, including technical details on data storage and security features of Jeeves. Ovulation test strips also had to be ordered for the study participants. Following the Christmas holidays, another meeting was held at the end of January to discuss the study's requirements, and the capabilities of Jeeves in fulfilling them. Deborah and Lucy watched the Jeeves tutorial video in order to understand the available features. Rather than beginning to design the study immediately after watching the video as before, there was a week gap during which Lucy planned her study design

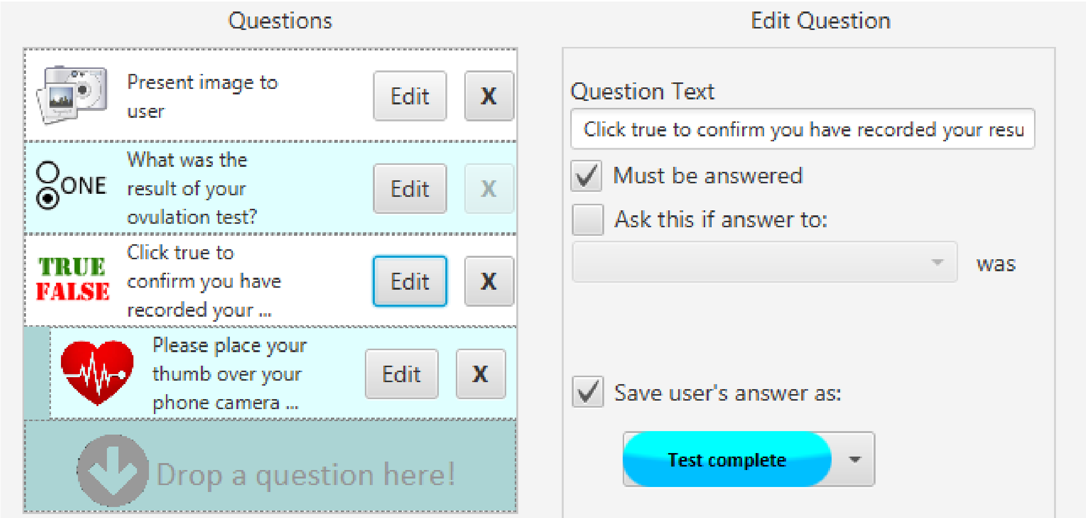


Figure 8.3
The survey created by Deborah and Lucy (with an attribute designating confirmation of completion)

around the present capabilities of Jeeves.

The research question again determined the study design. However, unlike the previous case study, the use of attributes was required in order to tailor the app to each participant's ovulation dates. Attribute creation appeared to be straightforward. Lucy created a survey question, created the date attribute, and then assigned the attribute to the question with no further issues, unlike the usability study participants in Chapter 6, who appeared to struggle with this sequence.

"you can create a survey that would get all the attributes out for the times so...then we can create an 'if this date send this survey' so that when it gets to the correct date they will just get another trigger" (Lucy)

An issue of *abstraction* occurred, whereby Lucy's study design required a date attribute, but also a second date attribute that represented 10 days after the first date. However, because date arithmetic was not possible, the researchers had to find a workaround. It was decided that, as the participant would be supervised by Lucy when completing this first questionnaire, a survey question could instruct participants to hand the device to Lucy so that she could enter the appropriate date manually, based on participants' previous response. A further issue of abstraction arose when the researchers desired to prompt a participant at a particular time if they had not completed their survey. A further workaround had to be employed, as indicated in the following dialogue:

Survey completion workaround

Deborah: I think then there must be some sort of 'this question has been answered' condition or 'this survey is complete' condition and if that's false, then...okay, this is probably not the best way to do it...but IF when we send the user input, when they do their 'please input your ovulation results' survey, we can add a question at the end that is just 'click here to confirm you've taken the test today' then you've got a yes/no answer to save it as an attribute again, which can then be placed into that condition.
Lucy: Okay. We could just put 'Click true to confirm you have recorded your results'. Oh, should I drag this [test complete attribute]?
Deborah: Yeah, 'if test complete is false'
Lucy: Oh no because they won't have selected false! Oh, is NOT true, but what if...only by selecting false it's not true, then they wouldn't have taken it, so it wouldn't come up...?

In summary, the researchers had to add a question to their survey that would update the "Test complete" attribute shown in Figure 8.2 to stop reminder prompts being sent. (Researchers' designed workaround survey is shown in Figure 8.3.) One further issue raised by this dialogue was that of default attribute values. Deborah, having had some programming experience, recognised that the Boolean attribute would likely default to be false. Although Lucy understood the concept of attributes, it was not previously considered that default variable values would be programmer-specific, tacit knowledge. Indeed, further programmer-specific knowledge was related to conditional expressions, specifically how often the condition would be checked, as made explicit in Deborah and Lucy's following discussion. This discussion related to a trigger they had created, as shown in Figure 8.4, where they were unsure whether the second prompt would be executed if the "Test complete" attribute had been updated to true:

Condition checking confusion

Deborah: As I understand it, it will send prompt, snooze for 60 minutes, send the prompt again, but will it check that condition again? So if the participant has then put true...
Lucy: I think so because it's underneath this one and you can put another one there...and also will it even have sent the initial prompt and then snoozed if it is true?
Deborah: If it's true then it won't do anything. Yeah. If it's false it'll prompt, wait, prompt again, but will it check, it doesn't look to me at the moment that it'll recheck the condition before it sends the second prompt
Lucy: Oh so if it did it the first time and then it snoozed it'd - so if the test complete is still false after 60 minutes, will it send it regardless? I don't think so...

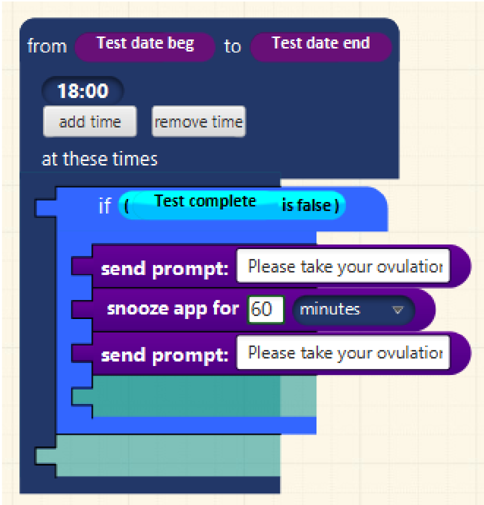


Figure 8.4
Researchers were unsure whether the second prompt would happen if the attribute was updated

Despite these minor problems, after 45 minutes, the researchers had finished designing their study, and expressed satisfaction that they had independently implemented the specification in this short time.

8.4.2.2 Conducting the pilot study

Following the direct observation of study implementation, a series of circumstances arose that prevented the pilot study actually being initiated until one month later. The specification was not viewed by the researchers during this time (as indicated by the "last accessed" date and time feature of Jeeves).

The pilot study was conducted in March, and ran for 21 days, during which participants were required to report their ovulation result daily for 10 days of the study. As this study was not conducted in-person as before, the study's progress relied exclusively on feedback from the researchers.

8.4.3 Discussion

This case study, previously defined as a "theoretical replication" of the first, had a number of differences that provided a different perspective on certain aspects of the practical use of Jeeves, while simultaneously aligning with other aspects. However, given the unavoidable delays prior to study commencement, and the time constraints of this research, it was not possible to conduct a semi-structured interview as before. This is a limitation of the work in this chapter, as it does

Table 8.4
Observations made in Deborah and Lucy's case study that relate to design factors in Chapter 6

Observation	Supporting Design Factor (Chapter 6)
Clarification was required on Boolean attributes and if-conditions	**Have modular tutorial videos**
Researchers wanted to create attributes based on other attribute values	**Support different levels of abstraction**
Some participants did not receive notifications at all scheduled times	**Bridge the gulf of evaluation**

not allow for a complete theoretical replication of the previous case study. This instead remains a priority for future work, as the full-scale study is currently in preparation.

Written feedback from the researchers was instead provided on their experience with conducting the pilot study, where they briefly reflected on the positive and negative features of Jeeves in doing so. This, in combination with their implementation observation, is sufficient to validate some design and utility factors derived in previous chapters.

8.4.3.1 Usability

During the preliminary observation, where Deborah and Lucy engaged in creating their study specification, it was encouraging to observe that they were able to conceptualise a study consisting of various triggers, conditional branching surveys, and attributes, with no further support beyond the brief tutorial video. Linking attributes to survey questions, and then using these attributes in the blocks-based specification, appeared to be straightforward for both researchers. Indeed, there were very few instances of usability issues during this initial implementation phase.

A non-programmer participant in a usability study of Chapter 6 suggested that Jeeves *"makes you do logic without realising you're doing it"* and that he did not have to *"think like a programmer"*. However, two issues were identified that directly related to assumptions about non-programmers' mental models, specifically the default values of attributes, and the time at which the if-condition is evaluated. In both these instances, the researchers had to directly seek clarification, because this information had not been included in the tutorial. These issues were not insurmountable *Design Barriers* as defined by Ko et al. [194], and instead were related specifically to Boolean attributes, and if-conditions. Thus, the design factor ***"have modular tutorial videos"*** could be employed here to describe particular features of individual blocks.

In addition, the design factor of ***"support different levels of abstraction"*** could also have been applied in this study. The limitations of Jeeves meant that date-based arithmetic could not be applied. However, it was suggested by these researchers that attribute values could be derived from other attribute values using such operations, similar to the function of spreadsheets that have cell values as functions of other cells.

An unprompted suggestion from Lucy during the direct observation session related to feedback from both the usability study and the previous case study, regarding the need to ***"bridge the gulf of evaluation"***. While the researchers completed their specification with very little assistance, they expressed a desire to preview the app's functionality in order to resolve any unexpected errors:

"I think the thing I probably want apart from this because I think it's gone well, is a preview possibly of what it looks like on the phone? I don't know how easy that would be though. It might just be easier just to get your phone up, you've got the app" (Lucy)

Indeed, during the pilot study, some participants failed to receive notifications for unknown reasons, such that it was unclear whether this was due to an error in JeevesAndroid, a fault in the study specification, or simply the constraints of a participants' particular device. Bridging the gulf of evaluation would have allowed more clarity with regards to this.

Regardless of the issues that arose for which the appropriate design factors have been restated, the visual blocks-based approach was observed to hold strong potential for use by non-programmer researchers. Both Deborah and Lucy acknowledged their lack of programming experience, but both were observed to begin discussing their design in terms of blocks terminology within minutes of having begun working with Jeeves. The relative complexity of their final study specification indicated an understanding of how Jeeves would function, albeit not necessarily of the underlying programmatic concepts themselves. Nevertheless, Deborah and Lucy's speed of adaptation suggests that other concepts such as loops could be tentatively integrated as blocks to broaden the capacity of Jeeves.

8.4.3.2 Usefulness

Following the previous case study, a concern arose that functionality for explicit context triggering, and individual participant tailoring, were not used. Despite their incidental absence in the ESM studies reviewed in Section 2.2, it was expected that provision of this functionality would influence its use. However, it was explained that their research question guided the functions used, and not the inverse. Nevertheless, the researchers in this study capitalised on these possibilities, creating multiple attributes and using them to configure different conditional functionality in the app itself. Although to some extent this validates the requirement that ***"Jeeves should allow experienced researchers to perform more complex tasks"***, the researchers in this study were not experienced; they desired complex functionality from the beginning.

While in the previous case study, researchers desired their data to be formatted for SPSS, the researchers in this study were satisfied with the Excel spreadsheet format:

"Reviewing participant information and downloading data was simple and straightforward. The transfer of data to an Excel document is really helpful and makes further possible data analysis easy to do"

This is a positive result, supportive of the factor that ***"Jeeves should integrate with existing research technology"***. However, it further emerged that the implemented heart rate function did not always give accurate readings. While commercial heart rate apps are available, their readings are often dependent on the participants' device amongst other factors. For future apps that require accurate readings, Jeeves should be able to integrate with more sophisticated heart rate monitoring technology.

In written feedback, the researchers identified another issue, unrelated to its functionality, but to participants' experience with Jeeves:

"Participants were unsure about the permanent notification that said 'Jeeves running' on their phone. This would show on their drop down menu and on their lock screen. Two participants dropped out of the study saying that the app was 'annoying' them due to this"

Interestingly, participants did not report that the survey notifications themselves were annoying, and the researchers expressed that *"notifications are effective and non-intrusive"*. Nor did they view the content of the app negatively. However, the presence of the Jeeves icon is a necessary feature in the current version of Android in order to ensure that background processes are not destroyed. It was surprising that the small icon would cause such irritation as to lead to study drop-out, validating the factor that ***"Jeeves should be non-intrusive in participants' lives***.

8.4.3.3 Meta-design

The role involved in this study as an outside observer was useful in confirming that Jeeves could be used almost completely independently of developer input. Indeed, only necessary technical support was contributed to this study, through implementation of a heart rate monitoring feature, and by clarifying uncertainties in the initial implementation phase. However, relinquishing the role of research collaborator resulted in necessary absence from many potentially useful discussions of study progress. Thus, this was less insightful in determining the process of

Table 8.5
Observations made in Deborah and Lucy's case study that relate to utility factors in Chapter 7

Observation	Supporting Utility Factor (Chapter 7) Jeeves should...
Researchers employed full range of Jeeves functionality	**...allow experienced researchers to perform more complex tasks**
Hastily implemented heart rate functionality did not work as expected	**...integrate with existing research technology**
Participants were lost due to "annoying" presence of the Jeeves taskbar icon	**...be non-intrusive in participants' lives**

psychology studies that was ascertained in the previous case study.

8.4.4 Summary

In summary, taking a different role in the meta-design approach in each study proved beneficial, and justified the time and resources involved in the theoretical replication of the previous study. From a research perspective, the role of "participant observer" was useful for exposing the full process of a research study. However, from a practical perspective, the capability of researchers to conduct this process independently, with only the Jeeves software and a brief tutorial video, is a promising result for the feasibility of its real-world application.

The primary design factor prevalent in both case studies relates to app quality - both pairs of researchers requested the ability to preview their app prior to deployment in order to ensure that unexpected errors would not occur. Quality with respect to app features was of mixed importance - the first researchers only made use of basic features, whereas the second researchers made full use of attributes and participant tailoring. In both cases, however, use of Jeeves was contingent on having extra functionality implemented, exemplifying the importance of a meta-design approach.

8.5 Clinical Observation

Interviews with clinicians in Chapter 7 revealed that practice nurses running chronic disease management clinics are potential stakeholders of Jeeves. It was also determined that perceived ease-of-use was contingent on enabling healthcare professionals to implement standard protocols as efficiently as possible. However, the university setting of these interviews was removed from the clinical context of interest, and as such clinicians did not have access to existing software or relevant materials that represented their practical processes. As an additional threat to external validity, the majority of clinicians interviewed were GPs who, as discussed in Chapter 3, are seldom involved in the management of chronic conditions, which are a primary use case for experience sampling in clinical practice. With respect to this point, Blomberg et al. argue for observation, as opposed to anecdotal evidence, for informing system design:

"The distinction between what people do and what they say is also related to the fact that people often don't have access to the inarticulated, tacit knowledge associated with certain activities" [209, p. 130]

Thus, in order to understand how chronic disease management clinics are conducted, and potential barriers to adoption of Jeeves in these settings, an observation session was arranged at North Glen Medical Practice in Glenrothes.

8.5.1 Research questions

As discussed at the beginning of this chapter, this observation partially addressed the research sub-question of *"What barriers do researchers and clinicians face that an EUD-ESM tool could alleviate?"* In doing so, the following questions were derived prior to the session, in order to structure notes taken during observations:

- How are chronic disease clinic appointments structured?
- What information is exchanged between patients and nurses?
- How do patients currently manage their conditions?
- What problems do they face in doing so?
- What, if any, is the role of technology in chronic disease management?

Establishing the barriers, benefits, and context of use of an EUD-ESM tool can only be partially achieved through interviews that are removed from this context itself. Thus, direct observation of chronic disease nurses in their work practice enabled contextual factors to be directly obtained, providing a broader perspective of the work setting than could be obtained through relying on clinicians' past experiences.

8.5.2 Process

The observation session was conducted over three hours, with one hour assigned to each of three members of clinical staff - a practice nurse running a diabetes clinic, another practice nurse running a hypertension clinic, and a GP interviewed in the previous phase of data collection.

Although GPs do not generally engage in chronic care during their appointments, this observation session was insightful in determining the contrasting structure of these appointments.

Within each hour, three patients were seen by each clinician. Audio recordings and computer use were not permitted within the clinical setting. Instead, notes were hand-written during the observation sessions, transcribed and expanded on within 24 hours of the recording process. The observation sessions were *naturalistic*, in that the observation aimed to disrupt the process as little as possible. As such, it was also not possible to ask questions during patients' appointment time, and instead brief discussions with clinicians between appointments were used to summarise the preceding appointment, and verify assumptions made in the note-taking process.

8.5.3 Diabetes management clinic

The observations made during the diabetes management clinic are divided into the following themes: appointment structure, information exchange, self-management, salient problems, and technology use. These are associated with the aforementioned sub-questions.

8.5.3.1 Structure

Each of the three appointments in this clinic were annual reviews of patients. During these reviews, patients were queried about their general health, general management of their diabetes, unusual events since their previous review, current medication, and any additional resources required. Informal discussion of lifestyle took place, medication and recent medical results were reviewed, and patients had opportunities to ask questions.

8.5.3.2 Information exchange

Where necessary, information provided by the nurse was educational in content. She explained that the most crucial aspect of appointments is to provide patients with knowledge to cope in emergency situations. Each patient was asked if they knew what to do in the event of a "hypo" (hypoglycemia is low blood glucose that occurs in diabetes mellitus patients). Patients are generally required to carry a source of sugar with them at all times, in addition to electronic glucose monitors. While all patients were fully aware of the necessary procedure, one patient explained that it was sometimes difficult to remember everything required of him.

Diabetic patients are burdened in the need to remember different aspects of their treatment regime. This memory burden is observed in annual review appointments, and the nurse commented that *"it's amazing the things people don't tell you about"* with respect to dangerous situations or symptoms that should be reported. Patients frequently forgot to inform nurses about emergency situations, which were only later discovered when the nurse viewed their recent medical history.

8.5.3.3 Self-management

The nurse explained that, contrary to the stereotype of older patients being less accepting of patient-centred care, diabetic patients of all ages are proactive in self-monitoring their conditions, and often enthusiastic in doing so. Patients are required to carry electronic glucose monitors with them, in order to ensure that their levels are safe prior to driving or engaging in other activities that would require concentration, which these patients accept as part of their management regime. As previously mentioned by clinicians in Chapter 7, patients often purchase their own monitors for home use (ranging from around £10-£50).

Indeed, one patient mentioned proactively monitoring his blood pressure, although this was not a necessary part of his self-management plan. While financial incentives are required to ensure compliance in ESM research studies, patients are intrinsically motivated by their own health. Compliance does not necessarily mean engagement, however. Another patient, while passively compliant with self-monitoring, explained that he was often unsure whether his monitored readings had any significance. The nurse further explained that clinicians try to avoid aggravating patients considered to be the “worried well” - patients whose conditions are stable and well-managed, but become fixated on quantification of their health, and are unable to live at ease without obsessive self-monitoring. Such patients increase the burden on clinicians through making unnecessary appointments for reassurance. For this reason, the nurse explained that there was no definitive self-management protocol that was generalisable to each patient. Rather, nurses were given flexibility to *“go with what people are comfy with”*.

8.5.3.4 Problems

Contrary to the expectation of non-compliance with self-monitoring as discerned from research studies, patients are highly compliant, but face barriers in terms of memory and information overload. Two of the three patients had their medication reviewed, revealing a considerable number of drugs for managing comorbidities present in these patients. The nurse required clarification from patients on medication they were currently taking, as medication lists are sometimes not synchronised across the practice. When these discrepancies are not checked, this can result in patients reordering unnecessary prescriptions.

Patients who manage these comorbidities face the most problems in the management of their healthcare. The nurse explained how they *“have a lot going on”*, including various appointments with different professionals, and a list of medicines that must be taken at different times. Indeed, one patient was not aware that he was due an annual review, and was surprised when he had received an appointment letter.

8.5.3.5 Technology

In terms of chronic disease management resources, patients were provided with a self-management plan on paper. The nurse explained that while patients were generally adherent and enthusiastic about having a personalised plan, the paper often got misplaced, so that the nurse would keep a copy herself to use for discussion with patients.

While expressing interest with regards to their results, patients did not have access to their personal information. Instead, the nurse was required to communicate this to patients through a form on her computer screen. Indeed, technology use was limited beyond the nurse's patient management application. She described how, as an optional service, some clinics send patients appointment reminders by text message, but that this is at the discretion of the clinic's management, and due to the logistics of implementation, is uncommon.

8.5.4 Hypertension management

The hypertension management clinic had a similar format to that of the previous clinic, involving a nurse with expertise in hypertension, who saw three patients for their annual reviews.

8.5.4.1 Structure

From discussion with the nurse, it emerged that annual reviews were a relatively new process established in the clinic. Patients had responded positively to the annual review process, and the practice had recently received a letter of praise from one patient describing the value of feedback obtained in these reviews. The nurse explained how patients managing chronic disease are more engaged with their treatment following feedback from a healthcare professional. Prior to annual reviews, patients would receive annual blood tests, but would not be directly given the results, and consequently would often fail to attend their blood test, as they were never given any form of useful feedback.

8.5.4.2 Information exchange

As with the diabetes clinic, there was a reliance on paper-based assessment forms and information leaflets for the transfer of information between patient and nurse. In particular, one patient had misplaced her assessment form for blood pressure readings prior to her clinic visit. She instead brought her readings on scraps of paper, of which two of the four had gone missing on her journey to the clinic. The nurse explained the laborious process of transcribing these readings onto a separate piece of paper, then scanning and uploading it into the patient's records.

Nevertheless, the clinic takes a patient-centred approach to education, with a variety of information leaflets. It was noted, however, that while leaflets provide generic disease information, and nurses can answer any specific questions, patients may not be aware of issues relevant to their disease. One patient had been sedentary because recent cold weather had affected her asthma symptoms. However, after wearing a scarf following advice in a news report, she found this significantly reduced her symptoms.

8.5.4.3 Self-management

Despite similar difficulties to the diabetic patients in independent management of their treatment, hypertension patients also appeared to be proactive in doing so. All three patients were satisfied with monitoring their blood pressure readings every day, expressed interest in the results, and engaged in active discussion with the nurse, rather than passively receiving information.

The nurse explained that patients frequently get *white coat hypertension*, described in Chapter 7 as the artificial raising of blood pressure readings caused by the unfamiliar clinical environment. To alleviate this, they are encouraged to take their blood pressure readings at home, in a familiar context that provides the most accurate results.

8.5.4.4 Problems

Forgetfulness, particularly in medication management, was a pertinent issue noted during observation in two older patients. As discussed in the diabetes clinic, patients would order repeat prescriptions of unnecessary medication, due to lack of communication between clinical staff. One patient was unsure of the medication she had at home, and what she was regularly taking. The nurse was required to go through this patient's list of medication, asking which were actually necessary.

Also identified in the diabetes clinic, the nurse explained that time pressure was the greatest burden on her working practice. She discussed that patients often have to be scheduled at the last minute, such that when patients fail to turn up to their scheduled appointments, it places a significant burden on her ability to keep the clinic running to this tight schedule. Although appointment reminders were sometimes made via phone calls, this placed additional responsibility on nurses on top of their standard schedule. Further, an interesting point was made that the elderly population with chronic disease may also suffer from deafness, such that phone reminders may not be effective.

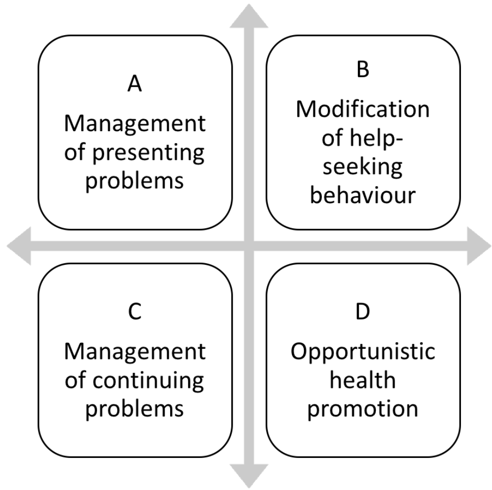


Figure 8.5
The four-point framework of general practice proposed by Stott and Davis [214, p. 201]

8.5.4.5 Technology use

The use of technology again appeared to be limited to the nurse's patient management application, which did not integrate well with patients' paper-based readings, as previously described in the manual transcription, scanning and uploading of patient results. The nurse enthusiastically described a new system for prescription management, whereby patients receive a text message to remind them that their prescription is due. The system has so far received a positive response from patients, but is limited to prescriptions and does not take into account regularly scheduled appointments, or reminders to take medication, for example.

8.5.5 GP clinic

Observation of the GP clinic was undertaken with the clinician identified as C1 in the interviews of Chapter 7. This clinician had previously explained the four-point framework of general practice, proposed by Stott and Davis [214], which is illustrated in Figure 8.5. In contrast to the chronic disease management clinics, the GP appointments were highly time-constrained, and involved directly focusing on the patients' acute problems, as illustrated in Section A of this model, leaving little or no time for the other sections related to chronic conditions, and holistic patient empowerment, to be addressed.

8.5.5.1 Structure

Although appointments are allocated to 10 minute slots, the first patient's appointment overran by an additional 15 minutes. Consequently, this delayed further scheduled appointments, which had to be caught up with. The GP explained that in many cases there simply is not an opportunity to address the underlying long-term health issues, when time has to be limited to discussion of acute issues. Nevertheless, patient-centred care was still a primary concern in the GP clinic, and patients' own interests and opinions were decisive in determining treatment. The GP explained:

"Patients' priorities are different from the doctor's priorities...You have to deal with this 'Presented Complaint'. There's no point if somebody comes in and says they've got a headache and you go 'yeah but we don't want to talk about that we want to talk about your blood pressure'."

8.5.5.2 Problems

Time constraints were particularly salient in the GP clinic, although also observed in the chronic disease management clinics to a lesser extent. This caused a burden on both the GP, and the patients themselves. Stott and Davis, in reference to their framework, state the following:

"It is too easy to treat each presenting problem but to fail to perceive the manipulations or cries for help which lie behind successive acute episodes of minor illness" [214, p. 203]

Although GPs acknowledge that dormant, chronic conditions are root causes of acute episodes, it is unreasonable to claim that their consultations are "too easy". The GP explained that patients often have recurring acute problems, but because of time issues, the symptoms are treated rather than the underlying cause. As a proxy to self-management advice, patients are sometimes directed to an informative website. However, the GP expressed doubt that this information is acted upon, stating: *"whether [website access] is followed up with or not is a different story."*

8.5.5.3 Technology

The GP described the system used to view patient information prior to appointments. He explained that it has been developed to prompt notifications of recommended actions to be taken for the selected patient. However, as previously discussed, although this information is useful, it cannot always be acted upon. The first patient had an acute problem that took significant time, but the system had notified the GP that he should also address her mental health. However, there was no time to do this after dealing with the acute issue. Nevertheless, the GP found this system useful in providing insights into patients that would otherwise require significantly more time manually searching their records.

In summary, the observation session revealed a number of implications for incorporation of Jeeves into clinical practice, that could not otherwise be obtained in context-independent settings. The following section describes these implications.

8.6 Implications for Design

Dourish argues that evaluating ethnography studies by their "implications for design" is a narrow perspective that fails to account for the analytic process involved in understanding users and their interactions with technology [215]. However, in this observation, the primary objective was simply to understand the barriers that an EUD-ESM tool could address, and therefore the requirements for the design of such a tool. Moreover, observations were not in-depth ethnographic studies, and instead served as a source of triangulation with clinician interviews. Table 8.6 summarises observations, their implications, and the supporting utility factors derived from interviews in Chapter 7, described in detail as follows.

8.6.1 Information viscosity

Information leaflets, self-management plans, forms for recording self-monitoring readings, and appointment reminders, were all paper-based. Given the prevalence of chronic disease in the elderly population, paper should be a universally accessible medium. However, it was interesting to note that paper-based communication often caused its own problems: patients would lose their self-management plans and recording forms, missed appointments were still an issue, and uploading paper-based readings into patient records was a laborious process.

In Chapter 7, clinicians suggested that the potential for Jeeves to integrate with existing patient management software would be useful. This is supported by observation in the hypertension clinic, where the nurse had to transcribe, scan and upload PDF documents into patients' health records, addressing the requirement that ***"Jeeves must be integrated with existing NHS technology"***. The laborious transfer from paper to electronic health records could be alleviated for nurses, but if patients have to do their own transfer of readings from glucose or blood pressure monitors into Jeeves, there still exists a labour on their part. Further, the requirement that ***"JeevesAndroid should support patients' current self-monitoring"*** could also streamline this process for patients themselves.

This information viscosity was also present between nurses and other clinicians. For example, changes to medication initiated by GPs were not immediately communicated to nurses through their system, such that nurses relied on patients to provide up-to-date information. Accessing results such as blood tests also appeared to be a time-consuming process through the nurses' system. This difficulty in communication was highlighted through interviews in Chapter 7, motivating the requirement that ***"Jeeves should make it easy for clinical staff to communicate and collaborate with each other"***.

8.6.2 Forgetfulness

The issue of forgetfulness is related to information viscosity - patients' difficulty in retrospective recall meant that potentially serious issues occurring in the previous year were not communicated at the time of their annual review. Chronic disease patients are expected to remember medication, self-monitoring equipment, appointments, as well as the adoption of lifestyle changes. This is further aggravated by the prevalence of comorbidities in diabetic and hypertension patients. The various aspects of their conditions that patients are required to remember means that a lot of responsibility is placed on them, with potentially life-threatening consequences of forgetfulness. For example, the diabetes nurse discussed the dangers of hypoglycemia when untreated with glucose; one hypertension patient mentioned running out of medication because she had forgotten to reorder in time.

Allowing patients and nurses to work together to schedule reminders on patients' devices for medication, appointments, and emergency equipment could be a highly useful feature, consistent with the factor that ***"Jeeves should support personalised reminders"***. As an example, a context-sensitive prompt could be issued to a patient when they leave their home to remind them to take appropriate equipment with them. This further links with the observation in Chapter 7 of the importance of app reliability. Ensuring that reminders are sent to patients consistently is of particular importance, such that ***"Clinicians must trust the resilience of developed apps"***.

8.6.3 Realistic adoption

As discussed in the previous chapter, the adoption of a tool like Jeeves must not impose additional burden on clinicians in its use. Although empowering patients to take responsibility for their healthcare would reduce the burden of unnecessary or missed appointments, it appears that the primary barrier to the adoption of Jeeves is the initial burden of introduction. Nurses are still highly constrained by time, and as such are unlikely to monitor patients' self-reported readings in between appointments unless specific time was allocated for them to do so. This supports the utility factor that ***"Clinicians must be allowed time to get used to Jeeves"***.

From the perspective of patients, realistic adoption is also contingent on a non-burdensome process. Traditional experience sampling studies require participants to complete surveys multiple times a day at random for a number of days or weeks. Direct application of ESM as used in psychology research is neither useful nor practical for patients managing chronic conditions. However, patients often have schedules that are defined by their medication. They may have medication to take in the mornings, afternoons, before or after meals, as well as less regular processes, such as monitoring their blood glucose or blood pressure. The diversity of patients' schedules and capabilities, as observed in chronic disease management clinics, is further evidence that ***"Jeeves should allow tailoring to individual patients"***.

The between-patient variation in skills and requirements necessitates individually tailored apps. However, *within*-patient differences of knowledge and proactive behaviour over time indicate the utility of allowing patients to tailor their apps independently. Recently diagnosed patients require considerable education and reminders to engage in treatment, but over time, their self-efficacy increases such that monitoring and medication become routine activities. Thus, continuous reminders from an app at this stage are likely to be irritating and intrusive. Allowing patients to adjust the level and type of reminders as necessary would therefore be beneficial, represented by the utility factor that ***"JeevesAndroid should be flexible to patients' needs***.

One limitation of these observations is that the inability to record interactions and directly ask patients and nurses questions during appointment times meant that the data collected and subsequent analysis do not fully represent the rich interactions that nurses have with patients, and with technology during the course of these appointments. Had permission been granted for audio or video recordings, or for the presence of a second observer to record notes, extra detail and a more thorough analysis could have been conducted. However, it is uncertain whether this would have significantly affected the course of the appointments. The purpose of these observations were to be as non-intrusive as possible, and patients were accepting of a single researcher with a notepad, whereas recording technology may have been intimidating.

8.7 Chapter Summary

This chapter described qualitative research that was conducted with an aim to give further validation to the factors derived in the previous two chapters. The case studies and observations addressed the lack of ecological validity in obtaining factors obtained from constrained usability studies, and from retrospective interviews in the offices of clinicians and researchers. The two case studies were particularly effective in determining whether usability factors in Chapter 6, and the utility factors in Chapter 7, would be validated in practical implementations of Jeeves.

Further, although it was not possible to deploy Jeeves in a clinical setting, naturalistic observation gave insights into the current working practices of chronic disease management nurses for whom Jeeves could potentially be useful. This served to validate some features of the interaction model derived in Chapter 3, and the utility factors of Chapter 7.

Table 8.6
Clinic observations and their relation to utility factors derived in Chapter 7

Observation	Patients are already self-monitoring, but have to bring their readings on paper
Implication	Allowing patients to synchronise the readings on their various electronic monitors would improve the process for patients and reduce errors in entry
Utility factor	**JeevesAndroid should support patients' current self-monitoring**
Observation	Viewing and updating patient records is a laborious process
Implication	Patient management software already exists, which Jeeves could integrate with to allow direct upload of patients' self-monitored readings, saving clinicians time.
Utility factor	**Jeeves must be integrated with existing NHS technology**
Observation	Patients' medication details and recent results are not always synchronised
Implication	A useful feature of any new technology in clinical practice would be allowing clinicians to immediately view updates made by another clinician
Utility factor	**Jeeves should make it easy for clinical staff to communicate and collaborate with each other**
Observation	Patients are burdened with keeping track of many responsibilities
Implication	The functionality of Jeeves to send time-contingent or context-sensitive reminder prompts could improve appointment turnout rates and medication adherence.
Utility factor	**Jeeves should support personalised reminders**
Observation	Failure to remember medication or monitoring equipment could be life-threatening
Implication	If clinicians develop reminders and prompts for patients, it is vital that these prompts arrive consistently if patients are to rely on them for their health
Utility factor	**Clinicians must trust the resilience of developed apps**
Observation	Clinical nurses have very little time in between appointments
Implication	If nurses are going to use Jeeves effectively to save time in the long run, they must be allowed to have scheduled time to monitor patients and track compliance
Utility factor	**Clinicians must be allowed time to get used to Jeeves**
Observation	Patients have a variety of comorbidities, requirements and personal schedules
Implication	A generic 'diabetes' or 'hypertension' app may not be appropriate for the diverse requirements of comorbid patients with individual struggles and demands
Utility factor	**Jeeves should allow tailoring to individual patients**
Observation	Considerable education must be given to patients before they become proactive
Implication	Jeeves could initially provide patients with information prompts and reminders to complete readings, the frequency of which patients could reduce as they become more proactive in self-monitoring.
Utility factor	**Jeeves should be flexible to patients' needs**

CHAPTER NINE

CONCLUSION AND FUTURE WORK

This chapter summarises the knowledge contribution made, by outlining the extent to which the original thesis objectives were achieved, their implications, as well as the limitations of this research and how future work could address these limitations.

9.1 Summary

The overarching goal of the work in this thesis was to enable psychologists and clinicians to create their own experience sampling apps for research participants or patients. The wealth of ESM studies in psychology and medical literature demonstrates the desire for these end-users to do so. However, in their reliance on professional programmers or existing creation tools, they must either sacrifice time, money, flexibility, functionality, or a combination thereof.

The introduction of this thesis defined a research question, its relevant sub-questions, and a set of objectives towards a solution to this identified problem. The primary question was outlined as follows:

Research Question

What are the factors influencing the adoption of technology for researchers and clinicians to develop experience sampling smartphone apps?

The literature review in Chapter 3 showed that many opportunities for smartphone ESM apps exist that could improve compliance, address more complex research questions, and support participants' well-being. However, Chapter 4 highlighted that currently available EUD tools do not adequately support these functions. While a professional programmer could supply this functionality, this approach is inflexible to changes required both during and after the study. Therefore, addressing this problem is important, because an appropriately designed EUD tool would allow researchers and clinicians to create context-sensitive, personalised ESM apps. Furthermore, it would allow them to adapt these apps as necessary for future studies, or even to cope with changing requirements during the course of a study. This thesis has presented the design and implementation of Jeeves, with a series of evaluations from which factors related to usability, usefulness, and organisational constraints have been derived.

9.1.1 Research sub-questions

A brief summary of the findings of this thesis follows, organised by the research sub-questions addressed. Following this, the contribution to knowledge that these findings provide will be discussed.

What difficulties do researchers and clinicians face that allowing them to develop experience sampling apps could alleviate?

This question was primarily addressed in Chapters 7 and 8, through interview and observation.

Psychology researchers are constrained by available research methods in investigating their phenomena of interest. As previously discussed, aggression research is limited by cross-sectional experiments that could not simulate the behaviour of interest; crowd research is limited by the difficulty of handing out and collecting paper surveys. Employing more sophisticated methods was hindered by a lack of time and money - such researchers do not have the time to learn new software, nor the money to have a professional programmer develop it for them, such that the functionality of existing technology dictates answerable research questions.

Clinicians in all areas, ranging from general practitioners to specialist ophthalmologists, face increasing time pressure, primarily due to high appointment numbers. Acute problems of patients recur due to poor general health management, leading to repeated appointments that could ultimately be reduced by empowering patients to self-manage. Customised ESM apps could support patients to take more responsibility for their health through timely reminders to self-monitor, which could also be completed and transmitted to their electronic health record through such an app. Further, appointment reminders would reduce the time and financial burden caused by those missed.

In summary, cross-sectional methods preclude researchers from accessing events of interest in participants' lives. Similarly, clinical appointments are largely cross-sectional in that they offer limited insights into individuals' everyday health management. While the quality of healthcare and research have scope to be improved, both researchers and clinicians lack the time to make the necessary changes.

What are the necessary features of apps that facilitate remote assessment of clinical patients or research participants in their natural environments?

The interaction model derived from reviewed literature in Section 3.2 showed how researchers, participants and apps should be able to interact in various ways that would benefit all stakeholders. Features such as allowing participants to tailor apps to their own individual requirements, directly contact researchers, and view feedback on their collected data, are seldom addressed in experience sampling studies. Further, allowing researchers to create apps that utilise context-awareness and tailored feedback based on participants' self-reports were found to be useful features for answering a wider variety of research questions. Nevertheless, recent studies reviewed in Section 2.2 show that, despite the predictions of Intille [57], many of such possibilities afforded by smartphone ESM have not been realised.

How can the development effort be reduced to allow rapid creation of experience sampling apps with no programming?

This question was addressed in Chapter 6, with theoretical contribution from Chapter 5.

The primary finding related to this question is that the blocks-based programming paradigm already supports a low level of perceived effort in creation of functional ESM apps by non-programmers. Participants were able to understand the event-driven programming paradigm, and even apply concepts such as variables and conditional statements. This shows that, even if these concepts are not fully understood, they can be successfully applied through example-based learning, therefore blocks-based programming can reduce end-user effort to apply programmatic concepts quickly and successfully. Further, it was found that it is particularly important to make use of a comprehensive tutorial that communicates concepts and actions through video. While this may appear to be a superficial finding, it is of particular importance for visual programming. When users are not visually guided in dragging and dropping blocks together, their *slips* lead to *mistakes* - semantic misunderstandings that take more effort to overcome. A final finding towards reducing development effort relates to abstraction - the blocks-based paradigm can allow end-users to communicate their abstraction requirements more easily. Contrary to the suggestion that participatory design and usability evaluations should not be confused [203], participants in the final usability study were able to express desirable functionality in terms of new blocks.

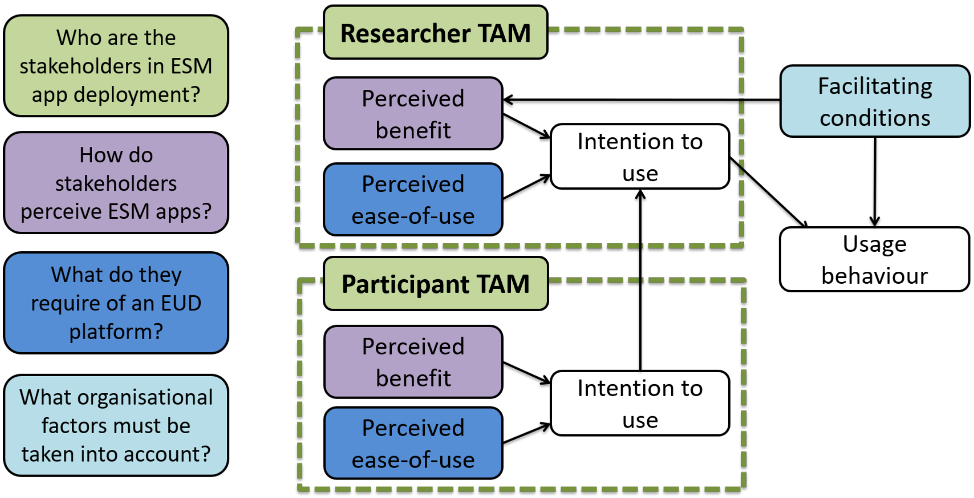


Figure 9.1
The ESM Technology Acceptance Model, illustrating the type of factors involved in adoption

How can researchers and clinicians employ an experience sampling app end-user development tool in practice?

This final question was addressed in Chapter 8 through case studies and observation.

A disappointing, albeit important finding, is that for clinicians to adopt Jeeves in practice, the main factor upon which this relies is not on ease-of-use nor usefulness, but on the *facilitating conditions* determined by their organisation. Clinicians are constrained by the health service's stringent requirements on app evaluation and ethical considerations, such that the *how can* question became a *why can't* question. In psychology, however, researchers felt strongly that they would like to use Jeeves in practice for in situ, longitudinal research. Further, the two case studies described in Chapter 8 document the processes by which two pairs of psychology researchers actually utilised Jeeves in realising their own research goals. The need arose for researchers to communicate and collaborate on studies for which they have their own particular areas of expertise. When it was not possible for researchers to work side-by-side on a project, they relied on chains of emails to collaborate. It was also found that providing complex functionality does not necessarily encourage its use - the extent to which researchers use Jeeves will likely be contingent on their research question. However, in contrast, the second pair of researchers generated new ideas from this availability of novel features, suggesting that the provision of complex functionality will encourage the generation of more ambitious research questions.

9.1.2 Implications

Synthesising the results from these sub-questions, the combined findings have implications for the critical adoption factors of EUD tools for experience sampling apps. The overarching themes of *time* and *quality* guide factors related to perceived ease-of-use, perceived usefulness, and facilitating conditions, as illustrated in Figure 9.1.

9.1.2.1 Time

Time appears to be the most critical barrier faced by researchers and clinicians, thus the time Jeeves would ultimately save (*perceived usefulness*) the time it would require to learn and use (*perceived ease-of-use*) and the time that organisations would allow for integration (*facilitating conditions*) are determining factors for adoption.

First, the time Jeeves would save is contingent on the specific goals of researchers and clinicians. For example, clinicians desired reminders to be used for appointments, and researchers desired a means to save time in acquiring informed consent from participants. When researchers used Jeeves in practice, time-saving qualities (such as a means to obtain informed consent, or collaboration features) emerged through direct use, and were not previously considered. This implies that a meta-design implementation, where end-users have a stake in design *during* use, is a necessary factor for sustained adoption. Both the case studies and usability studies demonstrated that end-users were able to articulate their time-saving requirements easily. While this cannot be conclusively attributed to the blocks-based paradigm, a key factor for enabling meta-design, and therefore adoption, is a representation that allows end-users to communicate their requirements effectively.

The blocks-based paradigm may not be the only means to support the rapid learnability of Jeeves. Section 5.3 describes how Jeeves is designed around the Cognitive Dimensions of Notations [161] and the Physics of Notations [162]. Further, features of Jeeves implemented with respect to these two frameworks were referenced positively by participants in usability and case studies. It follows that a visual language designed from the principles of these two frameworks supports learnability, and is therefore a factor for adoption. Although the Cognitive Dimensions framework is not prescriptive, the dimensions serve as important factors to consider in designing a learnable visual language. Section 6.5 has derived further guidelines from triangulating these frameworks with usability study results.

In terms of organisational constraints, time is also a *facilitating condition* for adoption in clinical practice. The National Health Service, or indeed any organisation that a clinician is part of, must allocate time for a period of adjustment to Jeeves. However, organisations' willingness to do so is contingent on the assurance that Jeeves will eventually save time, and therefore money. Empirical evidence of time-saving capabilities through an NHS-supported evaluation is thus necessary, and could be applied in future work.

9.1.2.2 Quality

Quality is another overarching factor discussed, which again can be considered in terms of the constructs in Figure 9.1. First, the quality of an app in terms of its functionality is a determining adoption factor (*perceived usefulness*), but particularly in terms of its reliability. A reliable app ensures that constant debugging and participant frustration are minimised (*perceived ease-of-use*), but is also necessary for organisations to ensure that apps will not potentially cause harm by malfunctioning (*facilitating conditions*).

Functional quality is critical for adopting new ESM technology. Section 7.3 showed how researchers are already comfortable with using *Qualtrics* software, which fulfills their needs with regards to survey creation. Clinicians in Sections 7.2 and 8.5 explained that target patients are often already active in self-monitoring. While this also implies that meta-design, where software can evolve to the needs of its users, is key to ensuring that apps are fit-for-purpose, researchers and clinicians also have initial requirements that must be satisfied by software which, as one researcher expressed it, *"let us do that which we couldn't otherwise do"*. While needs vary between end-users, the features derived from the interaction model in Section 3.2, namely: context-sensitivity, participant tailoring, automated feedback, and two-way communication, were considered desirable pre-requisites by interviewees in both domains.

Quality in terms of reliability is another primary adoption factor. Usability study participants in Chapter 6, and case study researchers in Chapter 8, expressed a strong desire to test their developed apps for correctness. Without this feature, frustration was caused by uncertainty as to whether their specifications satisfied the intended requirements. Direct collaboration with the case study researchers meant that personal assistance was given with app testing, but end-users need to be assured of app reliability prior to deployment if they are to consider adopting Jeeves independent of this firsthand support.

Reliability is another critical facilitating condition for organisations. Particularly in the health service, all apps must undergo rigorous evaluation to ensure that they will do no harm to patients. This is a difficult implication to address for an EUD tool, given that adoption is contingent on not just the reliability of one particular app, but on the reliability of all apps that could be developed by clinicians. Proof that clinicians themselves cannot implement potentially harmful apps with Jeeves is therefore a necessary adoption factor, and again would require an NHS-supported evaluation as future work.

9.1.3 Design guidelines

Summarising the implications derived from the analytical work of Chapters 2 and 3, and the empirical work of chapters 6 and 7, a concise set of design guidelines for the development of EUD tools for ESM apps is presented in this section. (Chapter 4 derived design decisions from existing EUD literature, but are not prescriptive guidelines in themselves, and Chapter 5 describes the implementation of these design decisions. Finally, the user research in Chapter 8 served to strengthen guidelines derived from previous chapters, thus new guidelines were not derived.) A detailed description of each can be found by consulting its respective chapter.

Guidelines from literature review (Chapters 2 & 3)

- **Incorporate implicit data collection from smartphone sensors** in order to infer interruptibility and minimise participant burden.
- **Incorporate prompting based on participants' unique responses** to both prompt for compliance, and to deliver emergency support messages.
- **Incorporate a channel of direct feedback between researcher and participant** for inevitable situations where direct human intervention is necessary.
- **Incorporate a means for participants to tailor apps to their personal preferences** to account for the diversity of study samples, and the participants within them.

Guidelines from usability studies (Chapter 6)

- **Employ blocks-based programming**, a paradigm which was seen as intuitive and effective in all three usability studies
- **Use text to distinguish graphics**, as users initially think in terms of nouns such as "survey" rather than semantics such as "trigger" or "action".
- **Minimise sequences of actions** where possible, to afford *recognition over recall* that is a key benefit of blocks-based programming
- **Make syntactic correctness easier** by highlighting or directing the user towards correct combinations.
- **Maximise visible blocks**, which relates to minimising action sequences to support recognition over recall.
- **Have 'tidying up' features**, analogous to code clean-up features in text-based IDEs, to improve the readability of specifications.
- **Support different levels of abstraction** for users at different levels of experience.
- **Have modular tutorial videos** that focus on commonly used, specific functions.
- **Allow reuse of examples** that more experienced users have previously created, to save time and foster learning by example.

- **Bridge the gulf of evaluation** to ensure that apps behave as users expect.

Guidelines from researcher and clinician feedback (Chapter 7)

- **Implement community support approaches** both within research groups, such as shared editing and annotations, and between groups, such as support forums.
- **Ensure participant data is encrypted with up-to-date security protocols**, otherwise ethical implications will prevent adoption.
- **Support integration with stakeholders' existing technology** - both researchers and clinicians have familiar technology that should be considered when introducing novel tools.

9.2 Contributions

The specific theoretical contributions of this thesis are:

- **A literature review of user-centred design studies and strategies for maximising ESM study utility, from which a model of useful ESM app features is derived (Chapter 3).** Primarily, this review and model contribute to psychology and medicine, by providing a set of features, derived from qualitative and quantitative evaluations in literature, that can improve engagement and compliance in ESM studies, and allow researchers and clinicians to assess participants and patients more effectively. These features were further found to be valuable in interviews and case studies.

- **A series of user studies that demonstrate the usability of the Jeeves blocks-based programming paradigm, with implications for domain-specific EUD and blocks-based programming (Chapter 6).** This particular contribution is to the end-user development community. By showing that end-users with no programming experience can create and modify complex ESM apps within a single one-hour usability evaluation, this demonstrates the applicability of blocks-based programming to domain-specific end-user development opportunities. In particular, the ability for participants to utilise programmatic constructs such as conditional statements and variables suggests that EUD tools in domains where such constructs could be useful, would benefit from a blocks-based approach.

- **A qualitative analysis of interviews, observations and case studies involving researchers and clinicians, from which a set of requirements are derived for ESM app EUD (Chapters 7 & 8).** This analysis and the derived requirements are a contribution to human-computer interaction, specifically to researchers developing software for use in the domains of psychology research or clinical practice. Regardless of whether software is intended to support end-user development, or simply to assist in end-users everyday work practices, the analysis of professionals in these two domains, their barriers, and consequent software requirements can guide future HCI developments.

- **A set of design guidelines for the development of EUD tools for ESM app creation (Chapter 9).** While these guidelines are presented earlier in this concluding chapter, they represent a synthesis of design implications derived from previous chapters, and are therefore a theoretical contribution spanning the entirety of this thesis.

Implementation of the Jeeves platform represents the overall practical contribution of this thesis. Moreover, its novel features represent individual contributions to EUD-ESM:

- **Meta-tailoring -** Work in this thesis has shown how ESM apps are deployed for a variety of purposes, and to different participant or patient groups. Thus, the level of flexibility that app end-users should have cannot be determined by meta-designers in advance. Jeeves, unlike other tools, provides a novel means of allowing researchers to tailor this level of participant flexibility, defined as *meta-tailoring*.

- **Context-sensitivity -** Jeeves lowers the barriers to context-sensitive ESM through allowing researchers to specify high-level concepts of location, activity, and device usage that can trigger further functions. Combined with meta-tailoring, researchers can also implement location-based triggers for semantic locations such as “Home” or “Work” that participants can define the exact locations of themselves.

- **Event-state interventions -** Finally, Jeeves allows researchers to define conditions under which to deliver tailored interventions. In distinguishing *events* (such as survey completion or arrival at a location) from explicitly reported *states* (such as mood or glucose level) researchers can combine these events and states to deliver in-the-moment, tailored intervention feedback to participants, previously impossible in other tools.

9.3 Limitations

Despite the potential of Jeeves to resolve a number of issues encountered in the design and deployment of ESM apps, some limitations exist in the research process undertaken in this thesis. While limitations also exist in the technical implementation of Jeeves and JeevesAndroid, these are discussed in the section on future work, as ideas for further design and implementation.

9.3.0.1 Late researcher involvement

The iterative development cycle of the Jeeves platform was largely focused on its usability by non-programmers, in that interviews with psychology researchers and clinicians only took place after a fully functional, usability-evaluated prototype had been developed. By not involving these researchers and clinicians at earlier stages of the Jeeves design, the development process was not user-centred from an early stage. Future iterations of Jeeves will ensure that feedback from expected end-users is incorporated from the beginning, and continuously applied to guide the development process. Nevertheless, the functional prototype shown in interviews acted as a means of both *creativity* and *communication*, as defined by Beaudouin-Lafon and Mackay [216]. Jeeves acted as a common ground for communication, where researchers and clinicians could conceptualise potential features in terms of the available blocks and survey questions. While this could have been performed with a low-fidelity prototype prior to usability evaluations, the previous research agenda was to have these researchers and clinicians conduct a similar usability study with Jeeves, such that by involving them in the development of prototypes, this would bias their understanding of the application. As these usability evaluations were not conducted, this was an unnecessary precaution.

9.3.0.2 Incomplete evaluation of JeevesAndroid

While lab simulations can be run on context-aware mobile apps, it is impossible to reliably anticipate the variation of users' context and how an app will respond in these situations. The inability to test deployed applications without initially conducting a pilot study is one limitation of Jeeves. Although the case studies demonstrated that JeevesAndroid can successfully run both simple and complex studies, these are two specific cases, hence a limitation of this research is that there was not field evaluation of many of the implemented features that the Jeeves EUD tool allows researchers to employ. As well as the extensions proposed in the section on future work, these extensions will also involve comprehensive field evaluation of their application in JeevesAndroid.

9.3.0.3 Lack of practice nurse interviews

Half of the interviewed clinicians in Section 7.2 were GPs, which at the time was not considered to be an issue. Recruitment of GPs was not intentional, and was instead a direct consequence of convenience sampling in the university's School of Medicine. It further emerged that GPs are minimally involved in the chronic disease management of patients, with which Jeeves could facilitate. Ideally, further interviews should have been conducted with practice nurses, complementary to the observation sessions. Nevertheless, all clinicians were well-informed on the duties of practice nurses and therefore aware of their potential motivations and constraints for employing Jeeves. Further work is required to elicit the firsthand feedback of practice nurses through interviews or focus groups.

9.4 Future Work

While the limitations discussed are in the research methods applied, much of the future work for Jeeves and JeevesAndroid is derived from the final usability study in Section 6.4, and from the interviews and case studies of Chapters 7 and 8 respectively. This future work addresses practical extensions to the Jeeves platform itself.

9.4.0.1 Comparative studies with existing tools

As discussed in Section 6.3, although a comparative study was not satisfactory within the constraints of a lab-based usability study, it remains important future work to conduct a longitudinal comparison of overall user experience between Jeeves and existing tools for ESM app creation. It remains to be ascertained whether the blocks-based paradigm is more effective, efficient and satisfying than, for example, the flowchart paradigm of *MovisensXS* or the programming-by-specification interface of *PACO*. Furthermore, as other tools have their own individual features, a comprehensive comparative study would determine which of these features would be potentially useful, and which would be ultimately perceived as redundant by researchers and clinicians.

9.4.0.2 From ESM to Ambulatory Assessment

Feedback from both psychology researchers and clinicians suggested the utility of extending the contexts can be sensed and reacted to. By focusing on context that can be acquired solely through the smartphone, this inevitably limits the physiological data that could be collected from participants through external devices. Today, exercise wristbands, blood pressure monitors, and glucose readers are becoming more discreet, and part of chronic disease patients' everyday lives. Their acceptance of this technology, and their willingness to self-monitor, supports providing a means of communication between the JeevesAndroid app and these external devices.

9.4.0.3 Visual feedback feature

The interaction model of Section 3.2 shows how allowing participants to view their own collected data could be an intrinsic motivation to engage with ESM apps. Visualisations were shown to increase compliance, and visual feedback was desired in user-centred design studies in health-related applications. At present, this is not possible in Jeeves, and remains as future work for improving compliance and satisfaction of participants. As only one study could be found that empirically evaluated the benefits of including such a feature, a further evaluation would also be of benefit.

9.4.0.4 Does it work on iPhone?

Cross-compatibility is a key concern for researchers, acknowledged in both literature, and by interviewees in Chapter 7. At the time of writing, iOS devices, while less widely used than Android devices, still have a significant market share. Given the importance of maximising sample size in research-based applications of ESM, an iOS version inclusive of all functions in JeevesAndroid is a particularly large, but potentially highly valuable area of future work.

9.4.0.5 Summary

As a project that aims to incorporate a meta-design framework, the future work of this project is continuously evolving, and will continue to be reassessed in response to requirements from its end-user developers, and participant end-users themselves. A key motivation for meta-design is that there will be no perfect software that satisfies the requirements of all its end-users. If perfection was ever achieved, it could not be sustainable; research goals and practice methods are continuously evolving, as are the technologies on which these goals and methods rely. Applying the terminology of meta-design, the future work goals listed above constitute a *reseeding* of Jeeves. Once this takes place, a further *evolutionary growth* phase follows, through incorporating additional end-user feedback [118]. This evolutionary approach ensures that Jeeves software remains *soft*, satisfying the requirements of its adoptive end-users.

Appendix A

Appendix-A ESM Survey Publications

The following pages tabulate the details of the 164 publications that were included in the survey in Chapter 2. First, SPSS output of the statistical tests referenced in the survey are provided, followed by the table of studies itself. As in the survey by van Berkel et al. [56], details include the year of publication, sample size (in the column labelled *N*), number of days, sampling strategy, device ownership, reward given, and compliance. In addition, the total number of signals given in the study, and the type of device used for assessment completion are provided.

Sampling strategies are given as a combination of S/I/E representing signal-contingent, interval-contingent, and event-contingent respectively. Sampling schedules with a * also sampled objective sensor data from participants, either through the smartphone or an external device.

Note that the reference to Kleiman et al. [51] is included twice in the table. This is because the publication reported two separate experience sampling studies, hence there are 165 unique ESM studies from 164 publications.

A.1 ESM Survey Statistics

Descriptive statistics of general ESM studies

	Participants	Days	Signals	Compliance	Signals/Day
N	165	165	165	141	160
Mean	88.91	15.039	-238.76	76.4583	5.9677882
Median	59.00	7.000	42.00	80.0000	5.0000000
Std. Deviation	103.958	23.1256	1733.306	14.41446	4.77085975
Minimum	1	1.0	-9999	26.52	.14286
Maximum	923	182.0	910	100.00	50.00000

Pearson Correlation between study duration and compliance in general ESM studies

		Days	Compliance
Days	Pearson Correlation	1	.062
	Sig. (2-tailed)		.467
	N	165	141
Compliance	Pearson Correlation	.062	1
	Sig. (2-tailed)	.467	
	N	141	141

Pearson Correlation between number of signals a day and compliance in general ESM studies

		Signals/Day	Compliance
Signals/Day	Pearson Correlation	1	-.280**
	Sig. (2-tailed)		.001
	N	160	137
Compliance	Pearson Correlation	-.280**	1
	Sig. (2-tailed)	.001	
	N	137	141

**. Correlation is significant at the 0.01 level (2-tailed).

A.2 ESM Survey Table

Reference	Year	N	Days	Signals	Sampling	Device	Own device	Rewards	Compliance
Chen et al. [1]	2017	30	7	49	S	Smartphone	Mixed	?	77.60
McQuoid et al. [2]	2018	17	30	90	S/E*	Smartphone	Yes	Levelled	?
Li & Lansford [3]	2017	184	7	7	I	Smartphone	Yes	Fixed	?
Yoo et al. [4]	2017	91	7	35	I	Mobile	Yes	?	63.90
Band et al. [5]	2017	23	6	60	S	Smartphone	No	?	65.00
Comulada et al. [6]	2017	42	182	728	S/I	Smartphone	No	?	?
Kirchner et al. [7]	2017	90	7	35	S	Smartphone	No	?	?
Chaudhury et al. [8]	2017	50	7	42	S	PDA	No	?	69.05
Toomey & Rudolph [9]	2017	42	5	15		Other	?	Raffle	94.13
Strauss et al. [10]	2018	56	6	24	S	PDA	No	Levelled	89.00
Andrewes et al. [11]	2017	107	6	36	S	Mobile	No	Fixed	51.56
Carson et al. [12]	2017	50	7	21	I	PDA	No	Levelled	?
Demirci & Bogen [13]	2017	61	56	?	E	Smartphone	Yes	Levelled	79.00
Carels et al. [14]	2017	51	14	28	S/E	Smartphone	Yes	Fixed	56.30
Catterson et al. [15]	2017	215	6	36	I	Smartphone	Yes	Fixed	83.35
Pos et al. [16]	2017	124	6	60	S	PDA	No	?	63.00
Wu [17]	2017	80	7	56	S	PDA	No	Levelled	71.00
Lyusin & Mohammed [18]	2018	26	14	42	I	Paper	?	Fixed	95.00
Weiss et al. [19]	2017	44	10	10	I	Other	?	?	90.00
Kimhy et al. [20]	2017	40	1.5	15	S*	PDA	No	?	89.80
Lindert et al. [21]	2018	35	7	70	S/I*	Smartphone	No	Fixed	80.00
Stieger et al. [22]	2017	246	21	42	S	Smartphone	Yes	Fixed+Raffle	?
Liddle et al. [23]	2017	40	3	?	S	Smartphone	Mixed	?	?
Van Voorhees et al. [24]	2018	117	14	?	S/E	PDA	No	Levelled	72.90
Kluge et al. [25]	2017	18	4	40	S*	Other	No	Levelled	?
Pearson et al. [26]	2017	5	14	84	S/E*	Smartphone	Yes	Levelled	?
O'Connor et al. [27]	2018	144	8	33	S	Smartphone	Mixed	Fixed	52.00
Benedek et al. [28]	2017	38	14	14	I	Smartphone	Yes	None	92.00
Martínez-Sierra et al. [29]	2018	1	13	13	E	Smartphone	Yes	None	100.00
Brose et al. [30]	2017	202	7	70	S	Smartphone	No	Levelled	88.00
Shoham et al. [31]	2017	82	21	52	S/E	Smartphone	Yes	Fixed	?
Burkhart [32]	2017	145	10	50	S	Mobile	Yes	Fixed	69.00
Verhagen et al. [33]	2017	64	6	60	S	Smartphone	Mixed	None	52.70
Fuller-Tyszkiewicz et al. [34]	2017	73	7	42	S	Smartphone	Yes	Fixed	46.50

Reference	Year	N	Days	Signals	Sampling	Device	Own device	Rewards	Compliance
Uink et al. [35]	2017	108	7	35	S/I	Smartphone	No	None	57.00
Fuller-Tyszkiewicz et al. [36]	2018	161	7	42	S	Smartphone	Yes	?	51.95
Prinsen et al. [37]	2018	136	7	56	I	Smartphone	Yes	Fixed+Raffle	88.10
Phillips et al. [38]	2018	56	14	42	S	Mobile	Yes	Fixed	88.50
Schneider et al. [39]	2017	235	6	60	S	PDA	No	?	69.00
Santa Maria et al. [40]	2017	66	21	105	S/I	Smartphone	No	Levelled	43.00
Webber et al. [41]	2017	40	28	?	S	Mobile	Mixed	Levelled	56.30
Santangelo et al. [42]	2017	46	4	48	I	Smartphone	No	Levelled	82.00
Livingston et al. [43]	2017	50	14	84	S	Smartphone	No	Fixed	68.02
Forman et al. [44]	2017	189	28	168	S/E	Smartphone	No	Levelled	82.00
Moran et al. [45]	2017	31	7	28	S	Smartphone	No	Levelled	80.00
Knell et al. [46]	2017	238	7	7	I*	Smartphone	No	Levelled	92.90
Schumacher et al. [47]	2018	91	14	84	S	Smartphone	No	Levelled	80.13
Zenk et al. [48]	2017	97	7	35	S*	Smartphone	No	Fixed	70.30
Mason et al. [49]	2017	50	14	84	S/E	PDA	No	Levelled	80.00
Sells et al. [50]	2017	119	7	28	S	PDA	No	Levelled	76.00
Kleiman et al. [51]	2017	54	28	112	S/I/E	Smartphone	Yes	Levelled	62.75
Kleiman et al. [51]	2017	36	7	28	S/I/E	Smartphone	Mixed	Levelled	62.00
Liao et al. [52]	2017	110	4	32	S*	Smartphone	No	?	82.00
MacIntyre [53]	2017	74	7	35	S/I	Smartphone	Yes	Levelled	80.00
Bos et al. [54]	2018	40	30	90	I	Smartphone	Yes	?	84.44
Lenaert et al. [55]	2017	17	6	60	S	PDA	No	?	71.00
Nittel et al. [56]	2018	32	6	60	S	Other	No	Fixed	?
Moore et al. [57]	2017	20	7	35	S	Smartphone	No	?	86.40
Moitra et al. [58]	2017	65	28	112	S/I	Mobile	No	Levelled	30.00
Shiyko et al. [59]	2017	126	10	42	S/I	Mobile	No	Levelled	81.50
Hébert et al. [60]	2017	181	28	28	I/E	Smartphone	Yes	Levelled	80.00
Smiley et al. [61]	2017	25	14	42	S	Mobile	Yes	Levelled	57.30
Derrick et al. [62]	2017	62	21	105	S/I	Smartphone	No	Levelled	77.50
Kukk & Akkermann [63]	2017	158	3	21	S	PDA	No	?	79.40
Karwowski et al. [64]	2017	74	6	30	S	Mobile	Yes	Fixed	48.00
Van Berkel et al. [65]	2017	24	21	80	I/E*	Smartphone	Yes	Fixed	26.52
Pihet et al. [66]	2017	26	14	14	I	Smartphone	No	None	88.60
Gorka et al. [67]	2017	257	7	35	S/E	PDA	No	?	?

Reference	Year	N	Days	Signals	Sampling	Device	Own device	Rewards	Compliance
Connolly [68]	2017	121	7	28	S	Smartphone	Yes	Levelled	82.00
Manwaring et al. [69]	2017	68	98	23	E	Mobile	Yes	Fixed	88.00
Cotter & Silvia [70]	2017	132	7	98	S	Smartphone	Mixed	Fixed+Raffle	34.04
Pos et al. [71]	2017	50	6	60	S	PDA	No	?	?
Ponnada et al. [72]	2017	10	28	168	S	Other	No	None	59.91
Farmer et al. [73]	2017	32	42	294	I/E	Smartphone	No	Levelled	?
Klippel et al. [74]	2017	150	6	60	S	PDA	No	?	?
Fazeli et al. [75]	2017	109	7	21	I	Smartphone	No	?	75.14
Lüdtke et al. [76]	2017	35	2	8	I	Smartphone	Yes	Fixed	?
Manasse [77]	2017	12	14	98	S/I/E	Smartphone	Yes	Levelled	86.60
Xu et al. [78]	2017	95	14	28	I	Mobile	Yes	Raffle	?
Matz-Costa et al. [79]	2017	30	5	30	I	Other	No	?	89.70
Prud'homme [80]	2017	131	8	48	I	Other	?	Levelled	96.60
Marquet et al. [81]	2018	74	7	21	I*	Smartphone	Yes	?	?
Allen et al. [82]	2017	46	84	84	S	Smartphone	Mixed	Levelled	84.46
Bolman et al. [83]	2017	50	7	70	S	Paper	No	?	60.00
King et al. [84]	2017	61	21	?	E	PDA	No	Levelled	83.61
Robinson et al. [85]	2017	56	14	56	S	PDA	No	Levelled	71.43
Bassi et al. [86]	2018	13	5	20	S	Paper	?	?	95.00
Dzubur et al. [87]	2018	404	7	31	S	Smartphone	Mixed	Fixed	80.00
Connolly & Alloy [88]	2017	121	7	28	S	Smartphone	Yes	Levelled	82.00
Schöndube et al. [89]	2017	63	20	80	S	Smartphone	No	Fixed	86.75
Lapid Pickman et al. [90]	2017	182	30	60	I	Smartphone	Yes	Fixed	86.00
Holland et al. [91]	2017	81	7	70	S	Smartphone	Yes	Levelled	83.00
Syrek et al. [92]	2017	334	1	10	I	Other	?	?	74.00
Short et al. [93]	2017	30	8	32	S	Smartphone	Yes	Levelled	80.00
Paolillo et al. [94]	2017	35	14	56	S/I	Smartphone	No	Levelled	89.50
Uink et al. [95]	2018	206	7	35	S/I	Smartphone	No	None	62.00
Sala et al. [96]	2017	129	7	28	S	Mobile	Yes	Fixed	66.00
Choi et al. [97]	2017	603	17	50	S	Smartphone	Yes	Fixed	80.53
Bieg et al. [98]	2017	141	14	14	E	Other	No	?	90.20
Vansteelandt et a. [99]	2017	32	8	80	S	PDA	No	None	63.00
Bodin et al. [100]	2017	149	1	24	I	PDA	No	?	?
Vinci et al. [101]	2017	391	32	128	S/E	PDA	No	Fixed	77.00

Reference	Year	N	Days	Signals	Sampling	Device	Own device	Rewards	Compliance
Gonda-Kotani et al. [102]	2017	2	87	348	I	PDA	No	Levelled	99.70
Eddington et al. [103]	2017	55	7	56	S	Mobile	Mixed	?	63.66
Silva et al. [104]	2018	99	7	56	S	Paper	No	None	73.63
Hager et al. [105]	2017	160	8	53	S*	PDA	No	?	52.00
Yuan et al. [106]	2017	33	7	35	S	Mobile	Yes	Fixed	94.63
Beute & de Kort [107]	2018	59	6	48	S	Smartphone	No	Fixed	80.00
Wonderlich et al. [108]	2017	16	14	98	S/I/E	Smartphone	Yes	Levelled	84.00
Marcusson-Clavertz et al. [109]	2017	45	5	50	S	PDA	No	Fixed	78.00
Ketonen et al. [110]	2018	55	14	70	I	Smartphone	No	?	88.20
Scala et al. [111]	2018	54	21	126	S/E	Smartphone	No	?	74.00
Powell et al. [112]	2017	76	4	24	S	PDA	No	Fixed	91.10
Bedard et al. [113]	2017	96	5	35	S*	Smartphone	Yes	Levelled	64.00
Westermann et al. [114]	2017	15	7	70		Smartphone	No	Fixed	78.00
Fritz et al. [115]	2017	97	7	28	I*	Smartphone	No	?	95.00
Freeman & Gottfredson [116]	2017	122	21	84	S	Mobile	Yes	Levelled	73.45
Scott et al. [117]	2018	43	182	910	S*	Smartphone	No	Fixed	81.00
Holt [118]	2018	41	7	70	S	PDA	No	Fixed	89.00
Hennig et al. [119]	2017	61	7	21	I*	Smartphone	No	Fixed	89.10
Rose et al. [120]	2017	31	14	84	S	Smartphone	Yes	?	70.00
Shrier et al. [121]	2007	67	7	40	S/E	PDA	No	Levelled	52.00
Gloster [122]	2006	35	7	28	I	PDA	No	Fixed+Raffle	90.90
Rydin-Gray [123]	2007	38	14	14	S/E	PDA	No	Fixed	82.00
Freedman et al. [124]	2006	30	14	112	S/E	Mobile	No	Levelled	67.00
Kimhy et al. [125]	2006	20	1	10	S	PDA	No	?	80.50
Buysse et al. [126]	2008	65	7	28	I	PDA	No	?	92.10
Hacker & Ferrans [127]	2007	20	6	18	I	Other	No	?	86.50
Tong et al. [128]	2007	118	1	17	I	PDA	No	Raffle	64.70
Rowan et al. [129]	2007	65	7	28	S/I/E	PDA	No	Levelled	69.10
Van Winkel et al. [130]	2008	56	6	72	S	Paper	N/A	?	59.72
Abela et al. [131]	2007	56	42	6	S	PDA	No	?	?
Sonnenschein et al. [132]	2007	42	14	98	S/I	PDA	No	?	84.20
Kashdan & Steger [133]	2007	97	21	21	I	Paper	N/A	?	96.02
Granholm et al. [134]	2008	54	7	28	I	PDA	No	Fixed	69.00
Kane et al. [135]	2007	124	7	56	S	PDA	No	Raffle	77.70

Reference	Year	N	Days	Signals	Sampling	Device	Own device	Rewards	Compliance
Hopper et al. [136]	2006	22	42	232	S/I	Other	No	?	62.40
Salvy et al. [137]	2008	20	7	42	S	PDA	No	Fixed	87.14
Shiffman et al. [138]	2007	214	16	90	S/E	PDA	No	Fixed	91.00
Hussong [139]	2007	85	28	84	S	Paper	N/A	Fixed	?
Kikuchi et al. [140]	2006	40	7	28	S/E	Other	No	?	96.00
Boseck et al. [141]	2007	14	14	42	I/E	PDA	No	Fixed	83.40
Gorely et al. [142]	2007	923	4	200	I	Paper	N/A	?	?
Bailes [143]	2006	11	7	42	S	Paper	N/A	?	90.00
Pieters et al. [144]	2006	2	7	63	S	Paper	N/A	?	73.00
Hogarth et al. [145]	2007	74	10	60	S	Paper	N/A	Fixed	?
Trull et al. [146]	2008	60	28	168	S	PDA	No	Fixed	87.50
Juslin et al. [147]	2008	32	14	98	S	PDA	No	Fixed	74.00
Moberly & Watkins [148]	2008	93	7	56	S	Paper	N/A	Fixed	77.00
Grühn et al. [149]	2008	114	7	35	S	PDA	No	Fixed	?
Graham [150]	2008	20	7	49	S	PDA	No	Fixed	66.00
Snir & Zohar [151]	2008	65	7	28	S	Paper	N/A	Fixed	93.80
Leahey & Crowther [152]	2008	105	5	30	S	PDA	No	Fixed	84.50
Putnam & McSweeney [153]	2008	13	7	35	S	PDA	No	?	74.95
Hsieh et al. [154]	2008	25	5	50	S	Other	Yes	Fixed	71.50
Gwaltney et al. [155]	2008	13	21	84	S/E	PDA	No	Fixed	75.00
Knouse et al. [156]	2008	206	7	56	S	PDA	No	Fixed+Raffle	74.30
Waters & Li [157]	2008	44	7	28	S/E	PDA	No	?	81.20
Hausenblas et al. [158]	2008	40	6	24	S/E	Paper	N/A	?	84.60
Galloway et al. [159]	2008	6	35	105	S	Mobile	No	Levelled	65.00
Husky et al. [160]	2008	5	10	50	S	Mobile	No	Levelled	96.00
Sestak [161]	2008	41	10	60	I	PDA	No	?	76.00
Ji et al. [162]	2008	160	10	60	S	Paper	N/A	Fixed	97.55
Shernoff et al. [163]	2008	191	7	35	S	Paper	N/A	?	94.30
Miller [164]	2008	82	7	56	S	PDA	No	Levelled	74.70

A.3 ESM Table References

[1] Y. W. Chen, A. C. Bundy, R. Cordier, Y. L. Chien, and S. L. Einfeld, “A cross-cultural exploration of the everyday social participation of individuals with autism spectrum disorders in Australia and Taiwan: an experience sampling study,” *Autism*, vol. 21, pp. 231–241, Feb 2017.

[2] J. McQuoid, J. Thrul, and P. Ling, “A geographically explicit ecological momentary assessment (GEMA) mixed method for understanding substance use,” *Social Science & Medicine*, vol. 202, pp. 89–98, 2018.

[3] J. J. Li and J. E. Lansford, “A smartphone-based ecological momentary assessment of parental behavioral consistency: associations with parental stress and child ADHD symptoms,” *Developmental psychology*, 2017.

[4] H. H. Yoo, S.-K. Jun, S. Y. Kim, and K. H. Park, “A study on the daily life experience of medical students using the experience sampling method.,” *International Journal of Contents*, vol. 13, no. 4, pp. 16–22, 2017.

[5] R. Band, C. Barrowclough, K. Caldwell, R. Emsley, and A. Wearden, “Activity patterns in response to symptoms in patients being treated for chronic fatigue syndrome: an experience sampling methodology study,” *Health Psychology*, vol. 36, no. 3, pp. 264–269, 2017.

[6] W. S. Comulada, D. Swendeman, M. K. Koussa, D. Mindry, M. Medich, D. Estrin, N. Mercer, and N. Ramanathan, “Adherence to self-monitoring healthy lifestyle behaviours through mobile phone-based ecological momentary assessments and photographic food records over 6 months in mostly ethnic minority mothers,” *Public Health Nutrition*, vol. 21, no. 4, pp. 679–688, 2018.

[7] T. Kirchner, E. Magallón-Neri, M. S. Ortiz, I. Planellas, M. Forns, and C. Calderón, “Adolescents’ daily perception of internalizing emotional states by means of smartphone-based ecological momentary assessment,” *The Spanish Journal of Psychology*, vol. 20, p. E77, 2017.

[8] S. R. Chaudhury, H. Galfalvy, E. Biggs, T.-H. Choo, J. J. Mann, and B. Stanley, “Affect in response to stressors and coping strategies: an ecological momentary assessment study of borderline personality disorder,” *Borderline Personality Disorder and Emotion Dysregulation*, vol. 4, p. 8, Dec 2017.

[9] E. C. Toomey and C. W. Rudolph, “Age-conditional effects in the affective arousal, empathy, and emotional labor linkage: Within-person evidence from an experience sampling study,” *Work, Aging and Retirement*, vol. 4, no. 2, pp. 145–160, 2018.

[10] K. F. Visser, F. Z. Esfahlani, H. Sayama, and G. P. Strauss, “An ecological momentary assessment evaluation of emotion regulation abnormalities in schizophrenia.,” *Psychological medicine*, pp. 1–9, Jan 2018.

[11] H. E. Andrewes, C. Hulbert, S. M. Cotton, J. Betts, and A. M. Chanen, “An ecological momentary assessment investigation of complex and conflicting emotions in youth with borderline personality disorder,” *Psychiatry Research*, vol. 252, pp. 102–110, Jun 2017.

[12] R. L. Carson, J. J. Baumgartner, C. L. Ota, A. P. Kuhn, and A. Durr, “An ecological momentary assessment of burnout, rejuvenation strategies, job satisfaction, and quitting intentions in childcare teachers,” *Early Childhood Education Journal*, vol. 45, pp. 801–808, Nov 2017.

[13] J. R. Demirci and D. L. Bogen, “An ecological momentary assessment of primiparous women’s breastfeeding behavior and problems from birth to 8 weeks,” *Journal of Human Lactation*, vol. 33, pp. 285–295, May 2017.

[14] R. A. Carels, J. Rossi, C. Solar, and J. C. Selensky, “An ecological momentary assessment of weight stigma among weight loss participants,” *Journal of Health Psychology*, Feb 2017.

[15] A. D. Catterson, L. Eldesouky, and O. P. John, “An experience sampling approach to emotion regulation: situational suppression use and social hierarchy,” *Journal of Research in Personality*, vol. 69, pp. 33–43, Aug 2017.

[16] K. Pos, I. E. de Wit, F. A. van Dijk, A. A. Bartels-Velthuis, R. Bruggeman, C. J. Meijer, L. de Haan, B. Z. Alizadeh, N. J. Van Beveren, W. Cahn, P. Delespaul, I. Myin-Germeys, R. S. Kahn, F. Schirmbeck, C. J. Simons, N. E. van Haren, J. van Os, and R. van Winkel, “An experience sampling study on the ecological validity of the SWN-20: indication that subjective well-being is associated with momentary affective states above and beyond psychosis susceptibility,” *Psychiatry Research*, vol. 258, pp. 234–238, Dec 2017.

[17] H. Wu, J. Mata, D. J. Furman, A. J. Whitmer, I. H. Gotlib, and R. J. Thompson, “Anticipatory and consummatory pleasure and displeasure in major depressive disorder: an experience sampling study.,” *Journal of abnormal psychology*, vol. 126, no. 2, pp. 149–159, 2017.

[18] D. Lyusin and A.-R. Mohammed, “Are emotionally intelligent people more emotionally stable? An experience sampling study,” Tech. Rep. WP BRP 88/PSY/2018, National Research University Higher School of Economics (HSE), Moscow, Russia, Jan 2018.

[19] M. Weiss, S. Razinskas, J. Backmann, and M. Hoegl, “Authentic leadership and leaders’ mental well-being: An experience sampling study,” *The Leadership Quarterly*, vol. 29, no. 2, pp. 309 – 321, 2018.

[20] D. Kimhy, M. M. Wall, M. C. Hansen, J. Vakhrusheva, C. J. Choi, P. Delespaul, N. Tarrier, R. P. Sloan, and D. Malaspina, “Autonomic regulation and auditory hallucinations in individuals with schizophrenia: an experience sampling study,” *Schizophrenia Bulletin*, vol. 43, pp. 754–763, Jul 2017.

[21] B. H. W. te Lindert, J. Itzhacki, W. P. van der Meijden, M. L. Kringelbach, J. Mendoza, and E. J. W. Van Someren, “Bright environmental light ameliorates deficient subjective ‘liking’ in insomnia: an experience sampling study,” *Sleep*, vol. 41, no. 4, 2018.

[22] S. Stieger, D. Lewetz, and U.-D. Reips, “Can smartphones be used to bring computer-based tasks from the lab to the field? A mobile experience-sampling method study about the pace of life,” *Behavior Research Methods*, pp. 1–9, 2017.

[23] J. Liddle, A. Wishink, L. Springfield, L. Gustafsson, D. Ireland, and P. Silburn, "Can smartphones measure momentary quality of life and participation? A proof of concept using experience sampling surveys with university students," *Australian Occupational Therapy Journal*, vol. 64, pp. 294–304, Aug 2017.

[24] E. E. Van Voorhees, P. A. Dennis, E. B. Elbogen, B. Fuemmeler, L. C. Neal, P. S. Calhoun, and J. C. Beckham, "Characterizing anger-related affect in individuals with posttraumatic stress disorder using ecological momentary assessment," *Psychiatry Research*, vol. 261, pp. 274–280, Mar 2018.

[25] A. Kluge, M. Kirschner, O. M. Hager, M. Bischof, B. Habermeyer, E. Seifritz, S. Walther, and S. Kaiser, "Combining actigraphy, ecological momentary assessment and neuroimaging to study apathy in patients with schizophrenia," *Schizophrenia Research*, vol. 195, pp. 176–182, 2018.

[26] J. L. Pearson, H. Elmasry, B. Das, S. L. Smiley, L. F. Rubin, T. DeAtley, E. Harvey, Y. Zhou, R. Niaura, and D. B. Abrams, "Comparison of ecological momentary assessment versus direct measurement of e-cigarette use with a Bluetooth-enabled e-cigarette: a pilot study," *JMIR Research Protocols*, vol. 6, p. e84, May 2017.

[27] S. G. O'Connor, W. Ke, E. Dzubur, S. Schembre, and G. F. Dunton, "Concordance and predictors of concordance of children's dietary intake as reported via ecological momentary assessment and 24 h recall," *Public Health Nutrition*, vol. 21, no. 6, pp. 1019–1027, 2018.

[28] M. Benedek, E. Jauk, K. Kerschenbauer, R. Anderwald, and L. Grond, "Creating art: an experience sampling study in the domain of moving image art," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 11, pp. 325–334, Aug 2017.

[29] G. Martínez-Sierra, Y. Arellano-García, A. Hernández-Moreno, and C. Nava-Guzmán, "Daily emotional experiences of a high school mathematics teacher in the classroom: a qualitative experience-sampling method," *International Journal of Science and Mathematics Education*, pp. 1–21, Feb 2018.

[30] A. Brose, M. Wichers, and P. Kuppens, "Daily stressful experiences precede but do not succeed depressive symptoms: results from a longitudinal experience sampling study," *Journal of Social and Clinical Psychology*, vol. 36, pp. 196–220, Mar 2017.

[31] A. Shoham, P. Goldstein, R. Oren, D. Spivak, and A. Bernstein, "Decentering in the process of cultivating mindfulness: an experience-sampling study in time and context," *Journal of Consulting and Clinical Psychology*, vol. 85, no. 2, pp. 123–134, 2017.

[32] M. L. Burkhart, *Deconstructing the Paradox of Parenthood: Using Experience Sampling Methodology and Attachment Theory to Understand Mothers' Emotional Experiences*. PhD thesis, The Claremont Graduate University, Claremont, CA, USA, 2017.

[33] S. J. W. Verhagen, J. A. Berben, C. Leue, A. Marsman, P. A. E. G. Delespaul, J. van Os, and R. Lousberg, "Demonstrating the reliability of transdiagnostic mhealth routine outcome monitoring in mental health services using experience sampling technology," *PLOS ONE*, vol. 12, pp. 1–14, Oct 2017.

[34] M. Fuller-Tyszkiewicz, T. Karvounis, R. Pemberton, L. Hartley-Clark, and B. Richardson, "Determinants of depressive mood states in everyday life: an experience sampling study," *Motivation and Emotion*, vol. 41, pp. 510–521, Aug 2017.

[35] B. N. Uink, K. L. Modecki, and B. L. Barber, "Disadvantaged youth report less negative emotion to minor stressors when with peers: an experience sampling study," *International Journal of Behavioral Development*, vol. 41, pp. 41–51, Jan 2017.

[36] M. Fuller-Tyszkiewicz, B. Richardson, V. Lewis, J. Smyth, and I. Krug, "Do women with greater trait body dissatisfaction experience body dissatisfaction states differently? An experience sampling study," *Body Image*, vol. 25, pp. 1–8, Jun 2018.

[37] S. Prinsen, C. Evers, L. Wijngaards, R. van Vliet, and D. de Ridder, "Does self-licensing benefit self-regulation over time? an ecological momentary assessment study of food temptations," *Personality and Social Psychology Bulletin*, vol. 44, no. 6, pp. 914–927, 2018.

[38] K. T. Phillips, M. M. Phillips, T. L. Lalonde, and M. A. Prince, "Does social context matter? an ecological momentary assessment study of marijuana use among college students," *Addictive Behaviors*, vol. 83, pp. 154–159, 2018.

[39] M. Schneider, U. Reininghaus, M. van Nierop, M. Janssens, and I. Myin-Germeys, "Does the social functioning scale reflect real-life social functioning? an experience sampling study in patients with a non-affective psychotic disorder and healthy control individuals," *Psychological Medicine*, vol. 47, no. 16, pp. 2777–2786, 2017.

[40] D. S. Maria, N. Padhye, Y. Yang, K. Gallardo, G.-M. Santos, J. Jung, and M. Businelle, "Drug use patterns and predictors among homeless youth: Results of an ecological momentary assessment," *The American Journal of Drug and Alcohol Abuse*, vol. 44, no. 5, pp. 551–560, 2018.

[41] C. J. Webber, E. C. O'Hea, B. Abar, B. Bock, and E. D. Boudreaux, "Ecological momentary assessment and first smoking cessation lapse after an acute cardiac event: a pilot study," *Journal of Health Psychology*, Dec 2017.

[42] P. S. Santangelo, J. Koenig, V. Funke, P. Parzer, F. Resch, U. W. Ebner-Priemer, and M. Kaess, "Ecological momentary assessment of affective and interpersonal instability in adolescent non-suicidal self-injury," *Journal of Abnormal Child Psychology*, vol. 45, pp. 1429–1438, Oct 2017.

[43] N. A. Livingston, A. Flentje, N. C. Heck, A. Szalda-Petree, and B. N. Cochran, "Ecological momentary assessment of daily discrimination experiences and nicotine, alcohol, and drug use among sexual and gender minority individuals," *Journal of Consulting and Clinical Psychology*, vol. 85, pp. 1131–1143, Dec 2017.

[44] E. M. Forman, L. M. Schumacher, R. Crosby, S. M. Manasse, S. P. Goldstein, M. L. Butryn, E. P. Wyckoff, and J. Graham Thomas, "Ecological momentary assessment of dietary lapses across behavioral weight loss treatment: characteristics, predictors, and relationships with weight change," *Annals of Behavioral Medicine*, vol. 51, pp. 741–753, Oct 2017.

[45] E. K. Moran, A. J. Culbreth, and D. M. Barch, "Ecological momentary assessment of negative symptoms in schizophrenia: relationships to effort-based decision making and reinforcement learning," *Journal of Abnormal Psychology*, vol. 126, pp. 96–105, Jan 2017.

[46] G. Knell, K. Pettee Gabriel, M. S. Businelle, K. Shuval, D. W. Wetter, and D. E. Kendzor, "Ecological momentary assessment of physical activity: validation study," *Journal of Medical Internet Research*, vol. 19, p. e253, Jul 2017.

[47] L. M. Schumacher, G. J. Martin, S. P. Goldstein, S. M. Manasse, R. D. Crosby, M. L. Butryn, J. Lillis, and E. M. Forman, "Ecological momentary assessment of self-attitudes in response to dietary lapses," *Health Psychology*, vol. 37, pp. 148–152, Feb 2018.

[48] S. N. Zenk, I. Horoi, K. K. Jones, L. Finnegan, C. Corte, B. Riley, and J. Wilbur, "Environmental and personal correlates of physical activity and sedentary behavior in African American women: an ecological momentary assessment study," *Women & Health*, vol. 57, pp. 446–462, Apr 2017.

[49] T. B. Mason, C. R. Pacanowski, J. M. Lavender, R. D. Crosby, S. A. Wonderlich, S. G. Engel, J. E. Mitchell, S. J. Crow, and C. B. Peterson, "Evaluating the ecological validity of the dutch eating behavior questionnaire among obese adults using ecological momentary assessment," *Assessment*, Jul 2017.

[50] J. R. Sells, A. J. Waters, and R. R. MacLean, "Evaluating the influence of at-risk alcohol use on factors associated with smoking cessation: combining laboratory and ecological momentary assessment," *Drug and Alcohol Dependence*, vol. 179, pp. 267–270, Oct 2017.

[51] E. M. Kleiman, B. J. Turner, S. Fedor, E. E. Beale, J. C. Huffman, and M. K. Nock, "Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies," *Journal of Abnormal Psychology*, vol. 126, pp. 726–738, Aug 2017.

[52] Y. Liao, C. P. Chou, J. Huh, A. Leventhal, and G. Dunton, "Examining acute bi-directional relationships between affect, physical feeling states, and physical activity in free-living situations using electronic ecological momentary assessment," *Journal of Behavioral Medicine*, vol. 40, pp. 445–457, Jun 2017.

[53] R. MacIntyre, "Examining body-focused self-improvement and self-evaluation social comparisons on exercise in young adult women using ecological momentary assessment," Master's thesis, Old Dominion University, Norfolk, VA, USA, Oct 2017.

[54] F. M. Bos, F. J. Blaauw, E. Snippe, L. van der Krieke, P. de Jonge, and M. Wichers, "Exploring the emotional dynamics of subclinically depressed individuals with and without anhedonia: an experience sampling study," *Journal of Affective Disorders*, vol. 228, pp. 186–193, Mar 2018.

[55] B. Lenaert, M. Colombi, C. van Heugten, S. Rasquin, Z. Kasanova, and R. Ponds, "Exploring the feasibility and usability of the experience sampling method to examine the daily lives of patients with acquired brain injury," *Neuropsychological Rehabilitation*, pp. 1–13, 2017.

[56] C. M. Nittel, T. M. Lincoln, F. Lamster, D. Leube, W. Rief, T. Kircher, and S. Mehl, "Expressive suppression is associated with state paranoia in psychosis: An experience sampling study on the association between adaptive and maladaptive emotion regulation strategies and paranoia," *British Journal of Clinical Psychology*, vol. 57, no. 3, pp. 291–312, 2018.

[57] R. C. Moore, C. N. Kaufmann, A. S. Rooney, D. J. Moore, L. T. Eyler, E. Granholm, S. P. Woods, J. Swendsen, R. K. Heaton, J. C. Scott, and C. A. Depp, "Feasibility and acceptability of ecological momentary assessment of daily functioning among older adults with HIV," *The American Journal of Geriatric Psychiatry*, vol. 25, pp. 829–840, Aug 2017.

[58] E. Moitra, B. A. Gaudiano, C. H. Davis, and D. Ben-Zeev, "Feasibility and acceptability of post-hospitalization ecological momentary assessment in patients with psychotic-spectrum disorders," *Comprehensive Psychiatry*, vol. 74, pp. 204–213, Apr 2017.

[59] M. P. Shiyko, S. Perkins, and L. Caldwell, "Feasibility and adherence paradigm to ecological momentary assessments in urban minority youth," *Psychological Assessment*, vol. 29, pp. 926–934, Jul 2017.

[60] E. T. Hébert, E. A. Vandewater, M. S. Businelle, M. B. Harrell, S. H. Kelder, and C. L. Perry, "Feasibility and reliability of a mobile tool to evaluate exposure to tobacco product marketing and messages using ecological momentary assessment," *Addictive Behaviors*, vol. 73, pp. 105–110, Oct 2017.

[61] S. L. Smiley, H. Elmasry, M. Webb Hooper, R. S. Niaura, A. B. Hamilton, and N. G. Milburn, "Feasibility of ecological momentary assessment of daily sexting and substance use among young adult african american gay and bisexual men: a pilot study," *JMIR Research Protocols*, vol. 6, p. e9, Feb 2017.

[62] J. L. Derrick, R. K. Eliseo-Arras, S. Haddad, M. Britton, and C. Hanny, "Feasibility of using ecological momentary assessment to study unaided smoking cessation in couples," *Nicotine & Tobacco Research*, pp. 1–10, 2017.

[63] K. Kukk and K. Akkermann, "Fluctuations in negative emotions predict binge eating both in women and men: an experience sampling study," *Eating Disorders*, vol. 25, pp. 65–79, Jan 2017.

[64] M. Karwowski, I. Lebuda, G. Szumski, and A. Firkowska-Mankiewicz, "From moment-to-moment to day-to-day: experience sampling and diary investigations in adults' everyday creativity," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 11, pp. 309–324, Aug 2017.

[65] N. Van Berkel, J. Goncalves, S. Hosio, and V. Kostakos, "Gamification of mobile experience sampling improves data quality and quantity," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, pp. 1–21, Sep 2017.

[66] S. Pihet, C. M. Passini, and M. Eicher, “Good and Bad Days: Fluctuations in the Burden of Informal Dementia Caregivers, an Experience Sampling Study,” *Nursing Research*, vol. 66, no. 6, pp. 421–431, 2017.

[67] S. M. Gorka, D. Hedeker, T. M. Piasecki, and R. Mermelstein, “Impact of alcohol use motives and internalizing symptoms on mood changes in response to drinking: an ecological momentary assessment investigation,” *Drug and Alcohol Dependence*, vol. 173, pp. 31–38, Apr 2017.

[68] S. L. Connolly, *Interplay Between Stress, Rumination, and Memory in Predicting Depression: an Ecological Momentary Assessment Study*. PhD thesis, Temple University, Philedelphia, PA, USA, 2017.

[69] K. C. Manwaring, R. Larsen, C. R. Graham, C. R. Henrie, and L. R. Halverson, “Investigating student engagement in blended learning settings using experience sampling and structural equation modeling,” *Internet and Higher Education*, vol. 35, pp. 21–33, Oct 2017.

[70] K. N. Cotter and P. J. Silvia, “Measuring mental music: comparing retrospective and experience sampling methods for assessing musical imagery,” *Psychology of Aesthetics, Creativity, and the Arts*, vol. 11, pp. 335–343, Aug 2017.

[71] K. Pos, C. J. Meijer, O. Verkerk, O. Ackema, L. Krabbendam, and L. de Haan, “Metacognitive training in patients recovering from a first psychosis: an experience sampling study testing treatment effects,” *European Archives of Psychiatry and Clinical Neuroscience*, vol. 268, pp. 57–64, Feb 2017.

[72] A. Ponnada, C. Haynes, D. Maniar, J. Manjourides, and S. Intille, “Microinteraction ecological momentary assessment response rates,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, pp. 1–16, Sep 2017.

[73] S. Farmer, D. Mindry, W. S. Comulada, and D. Swendeman, “Mobile phone ecological momentary assessment of daily stressors among people living with HIV: elucidating factors underlying health-related challenges in daily routines,” *Journal of the Association of Nurses in AIDS Care*, vol. 28, pp. 737–751, Sep 2017.

[74] A. Klippel, I. Myin-Germeys, U. Chavez-Baldini, K. J. Preacher, M. Kempton, L. Valmaggia, M. Calem, S. So, S. Beards, K. Hubbard, C. Gayer-Anderson, A. Onyejiaka, M. Wichers, P. McGuire, R. Murray, P. Garety, J. van Os, T. Wykes, C. Morgan, and U. Reininghaus, “Modeling the interplay between psychological processes and adverse, stressful contexts and experiences in pathways to psychosis: an experience sampling study,” *Schizophrenia Bulletin*, vol. 43, no. 2, pp. 302–315, 2017.

[75] P. L. Fazeli, J. M. Turan, H. Budhwani, W. Smith, J. L. Raper, M. J. Mugavero, and B. Turan, “Moment-to-moment within-person associations between acts of discrimination and internalized stigma in people living with HIV: an experience sampling study,” *Stigma and Health*, vol. 2, pp. 216–228, Aug 2017.

[76] T. Lüdtke, L. Kriston, J. Schröder, T. M. Lincoln, and S. Moritz, “Negative affect and a fluctuating jumping to conclusions bias predict subsequent paranoia in daily life: an online experience sampling study,” *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 56, pp. 106–112, Sep 2017.

[77] S. M. Manasse, *Novel Measurement of Affective Distress Intolerance: Behavioral Paradigm Development and Ecological Momentary Assessment in Individuals with Binge Eating*. PhD thesis, Drexel University, Philedelphia, PA, USA, 2017.

[78] W. Xu, X. Ding, Y. Zhuang, G. Yuan, Y. An, Z. Shi, and P. Hwa Goh, “Perceived haze, stress, and negative emotions: an ecological momentary assessment study of the affective responses to haze,” *Journal of Health Psychology*, Jul 2017.

[79] C. Matz-Costa, S. Cosner Berzin, M. Pitt-Catsouphes, and C. J. Halvorsen, “Perceptions of the meaningfulness of work among older social purpose workers: an ecological momentary assessment study,” *Journal of Applied Gerontology*, Aug 2017.

[80] J. Prud’homme, *Perfectionism and Daily Emotion Regulation and Well-being: An Experience Sampling Study*. PhD thesis, McGill University, Montréal, Québec, Canada, 2017.

[81] O. Marquet, C. Alberico, and A. J. Hipp, “Pokémon GO and physical activity among college students. A study using Ecological Momentary Assessment,” *Computers in Human Behavior*, vol. 81, pp. 215–222, Apr 2018.

[82] A. Allen, N. Tosun, S. Carlson, and S. Allen, “Postpartum changes in mood and smoking-related symptomatology: An ecological momentary assessment investigation,” *Nicotine & Tobacco Research*, vol. 20, no. 6, pp. 681–689, 2018.

[83] C. Bolman, P. Verboon, V. Thewissen, V. Boonen, K. Soons, and N. Jacobs, “Predicting smoking lapses in the first week of quitting: an ecological momentary assessment study,” *Journal of Addiction Medicine*, vol. 12, no. 1, pp. 65–71, 2018.

[84] E. B. King, J. J. Mohr, C. I. Peddie, K. P. Jones, and M. Kendra, “Predictors of identity management: an exploratory experience-sampling study of lesbian, gay, and bisexual workers,” *Journal of Management*, vol. 43, pp. 476–502, Feb 2017.

[85] C. D. Robinson, C. Muench, E. Brede, R. Endrighi, E. H. Szeto, J. R. Sells, J. P. Lammers, K. S. Okuyemi, G. Izmirlian, and A. J. Waters, “Pro-tobacco advertisement exposure among african american smokers: An ecological momentary assessment study,” *Addictive Behaviors*, vol. 83, pp. 142–147, 2018. Ambulatory Assessment of Addictive Disorders.

[86] M. Bassi, C. Rassiga, N. Fumagalli, and G. Senes, “Quality of experience during horticultural activities: an experience sampling pilot study among older adults living in a nursing home,” *Geriatric Nursing*, vol. 39, pp. 457–464, Jan 2018.

[87] E. Dzubur, J. Huh, J. P. Maher, S. S. Intille, and G. F. Dunton, “Response patterns and intra-dyadic factors related to compliance with ecological momentary assessment among mothers and children,” *Translational Behavioral Medicine*, vol. 8, no. 2, pp. 233–242, 2018.

[88] S. L. Connolly and L. B. Alloy, “Rumination interacts with life stress to predict depressive symptoms: an ecological momentary assessment study,” *Behaviour Research and Therapy*, vol. 97, pp. 86–95, Oct 2017.

[89] A. Schöndube, A. Bertrams, G. Sudeck, and R. Fuchs, "Self-control strength and physical exercise: an ecological momentary assessment study," *Psychology of Sport and Exercise*, vol. 29, pp. 19–26, Mar 2017.

[90] L. Lapid Pickman, T. Greene, and M. Gelkopf, "Sense of threat as a mediator of peritraumatic stress symptom development during wartime: an experience sampling study," *Journal of Traumatic Stress*, vol. 30, pp. 372–380, Aug 2017.

[91] E. Holland, P. Koval, M. Stratemeyer, F. Thomson, and N. Haslam, "Sexual objectification in women's daily lives: a smartphone ecological momentary assessment study," in *British Journal of Social Psychology*, vol. 56, pp. 314–333, Jun 2017.

[92] C. J. Syrek, J. Kühnel, T. Vahle-Hinz, and J. D. Bloom, "Share, like, twitter, and connect: Ecological momentary assessment to examine the relationship between non-work social media use at work and work engagement," *Work & Stress*, vol. 32, no. 3, pp. 209–227, 2018.

[93] N. A. Short, N. P. Allan, and N. B. Schmidt, "Sleep disturbance as a predictor of affective functioning and symptom severity among individuals with PTSD: an ecological momentary assessment study," *Behaviour Research and Therapy*, vol. 97, pp. 146–153, Oct 2017.

[94] E. W. Paolillo, L. C. Obermeit, B. Tang, C. A. Depp, F. Vaida, D. J. Moore, and R. C. Moore, "Smartphone-based ecological momentary assessment (EMA) of alcohol and cannabis use in older adults with and without HIV infection," *Addictive Behaviors*, vol. 83, pp. 102 – 108, 2018. Ambulatory Assessment of Addictive Disorders.

[95] B. Uink, K. L. Modecki, B. L. Barber, and H. M. Correia, "Socioeconomically disadvantaged adolescents with elevated externalizing symptoms show heightened emotion reactivity to daily stress: an experience sampling study," *Child Psychiatry & Human Development*, pp. 1–16, Feb 2018.

[96] M. Sala, L. C. Brosof, D. Rosenfield, K. C. Fernandez, and C. A. Levinson, "Stress is associated with exercise differently among individuals with higher and lower eating disorder symptoms: an ecological momentary assessment study," *International Journal of Eating Disorders*, vol. 50, pp. 1413–1420, Dec 2017.

[97] J. Choi, R. Catapano, and I. Choi, "Taking stock of happiness and meaning in everyday life: An experience sampling approach," *Social Psychological and Personality Science*, vol. 8, pp. 641–651, Aug 2017.

[98] M. Bieg, T. Goetz, F. Sticca, E. Brunner, E. Becker, V. Morger, and K. Hubbard, "Teaching methods and their impact on students' emotions in mathematics: an experience-sampling approach," *ZDM - Mathematics Education*, vol. 49, pp. 411–422, Jun 2017.

[99] K. Vansteelandt, M. Houben, L. Claes, A. Berens, E. Sleuwaegen, P. Sienaert, and P. Kuppens, "The affect stabilization function of nonsuicidal self injury in borderline personality disorder: an ecological momentary assessment study," *Behaviour Research and Therapy*, vol. 92, pp. 41–50, May 2017.

[100] F. Bodin, K. M. McIntyre, J. E. Schwartz, P. S. McKinley, C. Cardetti, P. A. Shapiro, E. Gorenstein, and R. P. Sloan, "The association of cigarette smoking with high-frequency heart rate variability: an ecological momentary assessment study," *Psychosomatic Medicine*, vol. 79, no. 9, pp. 1045–1050, 2017.

[101] C. Vinci, L. Li, C. Wu, C. Y. Lam, L. Guo, V. Correa-Fernández, C. A. Spears, D. S. Hoover, P. E. Etcheverry, and D. W. Wetter, "The association of positive emotion and first smoking lapse: an ecological momentary assessment study," *Health Psychology*, vol. 36, pp. 1038–1046, Nov 2017.

[102] C. Gonda-Kotani and G. W. White, "The effects of contingency contracts and performance feedback on completing data entries to self-monitor community participation of people with physical disabilities: an ecological momentary study," *Journal of Prevention and Intervention in the Community*, vol. 45, pp. 86–99, Apr 2017.

[103] K. M. Eddington, C. J. Burgin, P. J. Silvia, N. Fallah, C. Majestic, and T. R. Kwapil, "The effects of psychotherapy for major depressive disorder on daily mood and functioning: a longitudinal experience sampling study," *Cognitive Therapy and Research*, vol. 41, pp. 266–277, Apr 2017.

[104] E. Silva, T. Freire, and S. Faria, "The emotion regulation strategies of adolescents and their parents: An experience sampling study," *Journal of Child and Family Studies*, vol. 27, pp. 1774–1785, Jun 2018.

[105] E. R. Hager, N. A. Tilton, Y. Wang, N. C. Kapur, R. Arbaiza, B. C. Merry, and M. M. Black, "The home environment and toddler physical activity: an ecological momentary assessment study," *Pediatric Obesity*, vol. 12, pp. 1–9, Feb 2017.

[106] M. Z. Yuan, X. Guo, X. Li, X. jing Chen, C. Wang, and Y. Li, "The moderating role of regulatory emotional self-efficacy on smoking craving: An ecological momentary assessment study.," *PsyCh journal*, vol. 7, no. 1, pp. 5–12, 2018.

[107] F. Beute and Y. A. de Kort, "The natural context of wellbeing: ecological momentary assessment of the influence of nature and daylight on affect and stress for individuals with depression levels varying from none to clinical," *Health and Place*, vol. 49, pp. 7–18, Jan 2018.

[108] J. A. Wonderlich, L. E. Breithaupt, R. D. Crosby, J. C. Thompson, S. G. Engel, and S. Fischer, "The relation between craving and binge eating: integrating neuroimaging and ecological momentary assessment," *Appetite*, vol. 117, pp. 294–302, Oct 2017.

[109] D. Marcusson-Clavertz, S. Gušić, H. Bengtsson, H. Jacobsen, and E. Cardeña, "The relation of dissociation and mind wandering to unresolved/disorganized attachment: an experience sampling study," *Attachment and Human Development*, vol. 19, pp. 170–190, Mar 2017.

[110] E. E. Ketonen, J. Dietrich, J. Moeller, K. Salmela-Aro, and K. Lonka, "The role of daily autonomous and controlled educational goals in students' academic emotion states: an experience sampling method approach," *Learning and Instruction*, vol. 53, pp. 10–20, Feb 2018.

[111] J. W. Scala, K. N. Levy, B. N. Johnson, Y. Kivity, W. D. Ellison, A. L. Pincus, S. J. Wilson, and M. G. Newman, "The role of negative affect and self-concept clarity in predicting self-injurious urges in borderline personality disorder using ecological momentary assessment," *Journal of Personality Disorders*, vol. 32, pp. 36–57, 2018.

[112] D. J. Powell, C. Liossi, W. Schlotz, and R. Moss-Morris, "Tracking daily fatigue fluctuations in multiple sclerosis: ecological momentary assessment provides unique insights," *Journal of Behavioral Medicine*, vol. 40, pp. 772–783, Oct 2017.

[113] C. Bedard, S. King-Dowling, M. McDonald, G. Dunton, J. Cairney, and M. Kwan, "Understanding environmental and contextual influences of physical activity during first-year university: the feasibility of using ecological momentary assessment in the MovingU study," *JMIR Public Health and Surveillance*, vol. 3, p. e32, May 2017.

[114] S. Westermann, S. Grezellschak, Z. Oravecz, S. Moritz, T. Lüdtke, and A. Jansen, "Untangling the complex relationships between symptoms of schizophrenia and emotion dynamics in daily life: findings from an experience sampling pilot study," *Psychiatry Research*, vol. 257, pp. 514–518, Nov 2017.

[115] H. Fritz, W. Tarraf, D. J. Saleh, and M. P. Cutchin, "Using a smartphone-based ecological momentary assessment protocol with community dwelling older African Americans," *The Journals of Gerontology. Series B, Psychological Sciences and social Sciences*, vol. 72, pp. 876–887, Jan 2017.

[116] L. K. Freeman and N. C. Gottfredson, "Using ecological momentary assessment to assess the temporal relationship between sleep quality and cravings in individuals recovering from substance use disorders," *Addictive Behaviors*, vol. 83, pp. 95–101, 2018. Ambulatory Assessment of Addictive Disorders.

[117] C. K. Scott, M. L. Dennis, and D. H. Gustafson, "Using ecological momentary assessments to predict relapse after adult substance use treatment," *Addictive Behaviors*, vol. 82, pp. 72–78, Feb 2018.

[118] N. J. Holt, "Using the experience-sampling method to examine the psychological mechanisms by which participatory art improves wellbeing," *Perspectives in Public Health*, vol. 138, no. 1, pp. 55–65, 2018.

[119] T. Hennig, K. Krkovic, and T. M. Lincoln, "What predicts inattention in adolescents? An experience-sampling study comparing chronotype, subjective, and objective sleep parameters," *Sleep Medicine*, vol. 38, pp. 58–63, Oct 2017.

[120] S. W. Rose, A. Anesetti-Rothermel, H. Elmasry, and R. Niaura, "Young adult non-smokers' exposure to real-world tobacco marketing: results of an ecological momentary assessment pilot study," *BMC Research Notes*, vol. 10, p. 435, Dec 2017.

[121] L. A. Shrier, M. C. Shih, L. Hacker, and C. de Moor, "A momentary sampling study of the affective experience following coital events in adolescents," *Journal of Adolescent Health*, vol. 40, no. 4, pp. e1–e8, 2007.

[122] A. T. Gloster, *An ecological momentary assessment of retrospective memory accuracy in patients with obsessive-compulsive disorder*. PhD thesis, Eastern Michigan University, Ypsilanti, MI, USA, 2006.

[123] S. H. Rydin-Gray, *Binge eating antecedents among female college students: an ecological momentary assessment study*. PhD thesis, Bowling Green State University, Bowling Green, OH, USA, 2007.

[124] M. J. Freedman, K. M. Lester, C. McNamara, J. B. Milby, and J. E. Schumacher, "Cell phones for ecological momentary assessment with cocaine-addicted homeless patients in treatment," *Journal of Substance Abuse Treatment*, vol. 30, pp. 105–111, Mar 2006.

[125] D. Kimhy, P. Delespaul, C. Corcoran, H. Ahn, S. Yale, and D. Malaspina, "Computerized experience sampling method (ESMc): assessing feasibility and validity among individuals with schizophrenia," *Journal of Psychiatric Research*, vol. 40, pp. 221–230, Apr 2006.

[126] D. J. Buysse, W. Thompson, J. Scott, P. L. Franzen, A. Germain, M. L. Hall, D. E. Moul, E. A. Nofzinger, and D. J. Kupfer, "Daytime symptoms in primary insomnia: a prospective analysis using ecological momentary assessment," *Sleep Medicine*, vol. 8, pp. 198–208, Apr 2008.

[127] E. D. Hacker and C. E. Ferrans, "Ecological momentary assessment of fatigue in patients receiving intensive cancer therapy," *Journal of Pain and Symptom Management*, vol. 33, pp. 267–275, Mar 2007.

[128] E. M. W. Tong, G. D. Bishop, H. C. Enkelmann, Y. P. Why, S. M. Diong, M. Khader, and J. Ang, "Emotion and appraisal: a study using ecological momentary assessment," *Cognition and Emotion*, vol. 21, pp. 1361–1381, Nov 2007.

[129] P. J. Rowan, L. Cofta-Woerpel, C. A. Mazas, J. I. Vidrine, L. R. Reitzel, P. M. Cinciripini, and D. W. Wetter, "Evaluating reactivity to ecological momentary assessment during smoking cessation," *Experimental and Clinical Psychopharmacology*, vol. 15, no. 4, pp. 382–389, 2007.

[130] R. Van Winkel, C. Henquet, A. Rosa, S. Papiol, L. Fananás, M. De Hert, J. Peuskens, J. Van Os, and I. Myin-Germeys, "Evidence that the COMTVal158Met polymorphism moderates sensitivity to stress in psychosis: an experience-sampling study," *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, vol. 147, pp. 10–17, Jan 2008.

[131] J. R. Z. Abela, E. Morrison, and C. Starrs, "Excessive reassurance seeking, self-esteem, and depressive symptoms in children of affectively ill parents: an experience sampling analysis," *Journal of Social and Clinical Psychology*, vol. 26, pp. 849–869, Sep 2007.

[132] M. Sonnenschein, P. M. C. Mommersteeg, J. H. Houtveen, M. J. Sorbi, W. B. Schaufeli, and L. J. P. van Doornen, "Exhaustion and endocrine functioning in clinical burnout: an in-depth study using the experience sampling method," *Biological Psychology*, vol. 75, pp. 176–184, May 2007.

[133] T. B. Kashdan and M. F. Steger, “Expanding the topography of social anxiety: an experience-sampling assessment of positive emotions, positive events, and emotion suppression,” *Psychological Science*, vol. 17, no. 2, pp. 120–128, 2006.

[134] E. Granholm, C. Loh, and J. Swendsen, “Feasibility and validity of computerized ecological momentary assessment in schizophrenia,” *Schizophrenia Bulletin*, vol. 34, pp. 507–514, May 2008.

[135] M. J. Kane, L. H. Brown, J. C. McVay, P. J. Silvia, I. Myin-Germeys, and T. R. Kwapil, “For whom the mind wanders, and when: an experience-sampling study of working memory and executive control in daily life,” *Psychological Science*, vol. 18, pp. 614–621, Jul 2007.

[136] J. W. Hopper, Z. Su, A. R. Looby, E. T. Ryan, D. M. Penetar, C. M. Palmer, and S. E. Lukas, “Incidence and patterns of polydrug use and craving for ecstasy in regular ecstasy users: an ecological momentary assessment study,” *Drug and Alcohol Dependence*, vol. 85, pp. 221–235, Dec 2006.

[137] S. J. Salvy, J. W. Bowker, J. N. Roemmich, N. Romero, E. Kieffer, R. Paluch, and L. H. Epstein, “Peer influence on children’s physical activity: an experience sampling study,” *Journal of Pediatric Psychology*, vol. 33, pp. 39–49, Jan 2008.

[138] S. Shiffman, M. H. Balabanis, C. J. Gwaltney, J. A. Paty, M. Gnys, J. D. Kassel, M. Hickcox, and S. M. Paton, “Prediction of lapse from associations between smoking and situational antecedents assessed by ecological momentary assessment,” *Drug and Alcohol Dependence*, vol. 91, pp. 159–168, Dec 2007.

[139] A. M. Hussong, “Predictors of drinking immediacy following daily sadness: An application of survival analysis to experience sampling data,” *Addictive Behaviors*, vol. 32, pp. 1054–1065, May 2007.

[140] H. Kikuchi, K. Yoshiuchi, N. Miyasaka, K. Ohashi, Y. Yamamoto, H. Kumano, T. Kuboki, and A. Akabayashi, “Reliability of recalled self-report on headache intensity: investigation using ecological momentary assessment technique,” *Cephalalgia*, vol. 26, pp. 1335–1343, Nov 2006.

[141] J. J. Boseck, S. G. Engel, K. C. Allison, R. D. Crosby, J. E. Mitchell, and M. de Zwaan, “The application of ecological momentary assessment to the study of night eating,” *International Journal of Eating Disorders*, vol. 40, no. 3, pp. 271–276, 2007.

[142] T. Gorely, S. J. Marshall, S. J. H. Biddle, and N. Cameron, “The prevalence of leisure time sedentary behaviour and physical activity in adolescent girls: an ecological momentary assessment approach,” *International Journal of Pediatric Obesity*, vol. 2, pp. 227–234, Jan 2007.

[143] F. A. Bailes, “The use of experience-sampling methods to monitor musical imagery in everyday life,” *Musicae Scientiae*, vol. 10, pp. 173–190, Sep 2006.

[144] G. Pieters, K. Vansteelandt, L. Claes, M. Probst, I. Van Mechelen, and W. Vandereycken, “The usefulness of experience sampling in understanding the urge to move in anorexia nervosa,” *Acta Neuropsychiatrica*, vol. 18, pp. 30–37, Feb 2006.

[145] R. M. Hogarth, M. Portell, and A. Cuxart, “What risks do people perceive in everyday life? A perspective gained from the experience sampling method (ESM),” *Risk Analysis*, vol. 27, pp. 1427–1439, Dec 2007.

[146] T. J. Trull, M. B. Solhan, S. L. Tragesser, S. Jahng, P. K. Wood, T. M. Piasecki, and D. Watson, “Affective instability: measuring a core feature of borderline personality disorder with ecological momentary assessment,” *Journal of Abnormal Psychology*, vol. 117, no. 3, pp. 647–661, 2008.

[147] P. N. Juslin, S. Liljeström, D. Västfjäll, G. Barradas, and A. Silva, “An experience sampling study of emotional reactions to music: listener, music, and situation,” *Emotion*, vol. 8, no. 5, pp. 668–683, 2008.

[148] N. J. Moberly and E. R. Watkins, “Ruminative self-focus and negative affect: an experience sampling study,” *Journal of Abnormal Psychology*, vol. 117, pp. 314–323, May 2008.

[149] D. Grühn, K. Rebucal, M. Diehl, M. Lumley, and G. Labouvie-Vief, “Empathy across the adult lifespan: longitudinal and experience-sampling findings,” *Emotion*, vol. 8, no. 6, pp. 753–765, 2008.

[150] J. M. Graham, “Self-expansion and flow in couples’ momentary experiences: an experience sampling study,” *Journal of Personality and Social Psychology*, vol. 95, no. 3, pp. 679–694, 2008.

[151] R. Snir and D. Zohar, “Workaholism as discretionary time investment at work: an experience-sampling study,” *Applied Psychology*, vol. 57, pp. 109–127, Jan 2008.

[152] T. M. Leahey and J. H. Crowther, “An ecological momentary assessment of comparison target as a moderator of the effects of appearance-focused social comparisons,” *Body Image*, vol. 5, pp. 307–311, Sep 2008.

[153] K. M. Putnam and L. B. McSweeney, “Depressive symptoms and baseline prefrontal EEG alpha activity: a study utilizing ecological momentary assessment,” *Biological Psychology*, vol. 77, pp. 237–240, Feb 2008.

[154] G. Hsieh, I. Li, A. Dey, J. Forlizzi, and S. E. Hudson, “Using visualizations to increase compliance in experience sampling,” in *Proceedings of the 10th international conference on Ubiquitous computing*, (Seoul, Korea), pp. 164–167, ACM, 2008.

[155] C. Gwaltney, R. Bartolomei, S. Colby, and C. Kahler, “Ecological momentary assessment of adolescent smoking cessation: a feasibility study,” *Nicotine & Tobacco Research*, vol. 10, pp. 1185–1190, Jul 2008.

[156] L. E. Knouse, J. T. Mitchell, L. H. Brown, P. J. Silvia, M. J. Kane, I. Myin-Germeys, and T. R. Kwapil, “The expression of adult ADHD symptoms in daily life: an application of experience sampling methodology,” *Journal of Attention Disorders*, vol. 11, pp. 652–663, May 2008.

[157] A. J. Waters and Y. Li, "Evaluating the utility of administering a reaction time task in an ecological momentary assessment study," *Psychopharmacology*, vol. 197, pp. 25–35, Mar 2008.

[158] H. A. Hausenblas, L. Gauvin, D. S. Downs, and A. R. Duley, "Effects of abstinence from habitual involvement in regular exercise on feeling states: an ecological momentary assessment study," *British Journal of Health Psychology*, vol. 13, pp. 237–255, May 2008.

[159] G. P. Galloway, R. Didier, K. Garrison, and J. Mendelson, "Feasibility of ecological momentary assessment using cellular telephones in methamphetamine dependent subjects," *Substance Abuse*, vol. 1, pp. 9–14, Jan 2008.

[160] M. M. Husky, C. M. Mazure, K. M. Carroll, D. Barry, and N. M. Petry, "Using the experience sampling method in the context of contingency management for substance abuse treatment," *Journal of Applied Behavior Analysis*, vol. 41, pp. 635–644, Dec 2008.

[161] N. J. Sestak, *Psychological contagion within the supervisor-subordinate dyad: An experience sampling investigation of mood and job attitude contagion at work.* PhD thesis, University of Akron, Akron, OH, USA, 2008.

[162] L. Ji, C. Huet, and L. Dubé, "The emotional and social landscape of home vs away-from-home meals and their influence on eating pattern: an experience sampling study with non-obese adult women," in *Domestic Foodscapes: Towards Mindful Eating, Concordia University, Montreal*, 2008.

[163] D. J. Shernoff and D. L. Vandell, "Youth engagement in after-school programs: a perspective from experience sampling," in *Annual meeting of the American Educational Research Association, New York*, 2008.

[164] N. A. Miller, *Anxiety and emotion dysregulation in daily life: an experience-sampling comparison of social phobia and generalized anxiety disorder analogue groups.* PhD thesis, University of Nebraska, Lincoln, NE, USA, 2008.

APPENDIX B

APPENDIX-B ETHICAL APPROVAL DOCUMENTS

This Appendix contains the ethical approval letters that were received in order to conduct the usability evaluations, interviews and case studies in this thesis. The first letter confirmed ethical approval to conduct lab-based usability evaluations. The second letter gave permission to conduct case studies, interviews and observations with clinicians and psychology researchers. Finally, the third letter was required for conducting the pilot of the study involving experience sampling of sporting events with psychology researchers, as described in Chapter 7.

B.1 Lab Usability Studies

University of St Andrews
Scotland's first university – 1413

University Teaching and Research Ethics Committee Sub-committee

13th April 2015
Daniel Rough
School of Computer Science

Ethics Reference No: *Please quote this ref on all correspondence*	**CS11434**
Project Title:	"Rough Draft": Visual Programming for Mobile Applications
Researchers Name(s):	**Daniel Rough**
Supervisor(s):	**Professor Aaron Quigley**

Thank you for submitting your application which was considered at the Computer Science School Ethics Committee meeting on the 3rd April 2015. The following documents were reviewed:

1. Ethical Application Form 02/04/2015
2. Participant Information Sheet 02/04/2015
3. Coded data Consent Form 02/04/2015
4. Coded data Consent Form 2 02/04/2015
5. Debriefing Form 02/04/2015
6. Sample Advertisement 02/04/2015

The University Teaching and Research Ethics Committee (UTREC) approve this study from an ethical point of view. Please note that where approval is given by a School Ethics Committee that committee is part of UTREC and is delegated to act for UTREC.

Approval is given for three years. Projects, which have not commenced within two years of original approval, must be re-submitted to your School Ethics Committee.

You must inform your School Ethics Committee when the research has been completed. If you are unable to complete your research within the 3 three year validation period, you will be required to write to your School Ethics Committee and to UTREC (where approval was given by UTREC) to request an extension or you will need to re-apply.

Any serious adverse events or significant change which occurs in connection with this study and/or which may alter its ethical consideration must be reported immediately to the School Ethics Committee, and an Ethical Amendment Form submitted where appropriate.

Approval is given on the understanding that the 'Guidelines for Ethical Research Practice' https://www.st-andrews.ac.uk/utrec/guidelines/ are adhered to.

Yours sincerely

Convenor of the School Ethics Committee

Ccs Supervisor
School Ethics Committee

ethics-cs@st-andrews.ac.uk

The University of St Andrews is a charity registered in Scotland: No SC013532

B.2 Interviews and Observations

University of St Andrews | FOUNDED 1413

University Teaching and Research Ethics Committee

11 March 2018

Dear Daniel,

Thank you for submitting your ethical application which was considered at the School of Computer Science Ethics Committee meeting on 14th November when the following documents were reviewed:

1. Ethical Application Form
2. Participant Information Sheet
3. Consent Form
4. Debriefing Form

The School of Computer Science Ethics Committee has been delegated to act on behalf of the University Teaching and Research Ethics Committee (UTREC) and has granted this application ethical approval. The particulars relating to the approved project are as follows -

Approval Code:	CS13252	**Approved on:**	09.01.18	**Approval Expiry:**	09.01.23
Project Title:	Case Studies of Jeeves Visual Programming Platform				
Researcher(s):	Daniel Rough				
Supervisor(s):	Professor Aaron Quigley				

Approval is awarded for five years. Projects which have not commenced within two years of approval must be re-submitted for review by your School Ethics Committee. If you are unable to complete your research within the five year approval period, you are required to write to your School Ethics Committee Convener to request a discretionary extension of no greater than 6 months or to re-apply if directed to do so, and you should inform your School Ethics Committee when your project reaches completion.

If you make any changes to the project outlined in your approved ethical application form, you should inform your supervisor and seek advice on the ethical implications of those changes from the School Ethics Convener who may advise you to complete and submit an ethical amendment form for review.

Any adverse incident which occurs during the course of conducting your research must be reported immediately to the School Ethics Committee who will advise you on the appropriate action to be taken.

Approval is given on the understanding that you conduct your research as outlined in your application and in compliance with UTREC Guidelines and Policies (http://www.st-andrews.ac.uk/utrec/guidelinespolicies/). You are also advised to ensure that you procure and handle your research data within the provisions of the Data Provision Act 1998 and in accordance with any conditions of funding incumbent upon you.

Yours sincerely

Tristan Henderson

Convener of the School Ethics Committee

ethics-cs@st-andrews.ac.uk

The University of St Andrews is a charity registered in Scotland: No SC013532

B.3 ESM in Sporting Events Case Study

University of St Andrews | FOUNDED 1413

University Teaching and Research Ethics Committee

11 March 2018

Dear Daniel,

Thank you for submitting your amendment application which comprised the following documents:

1. Ethical Amendment Application Form
2. Participant Information Form
3. Advertisement
4. Questionnaire

The School of Computer Science Ethics Committee is delegated to act on behalf of the University Teaching and Research Ethics Committee (UTREC) and has approved this ethical amendment application. The particulars of this approval are as follows –

Original Approval Code:	CS13251	Approved on:	21.09.2017
Amendment Approval Date:	09.01.08	Approval Expiry Date:	21.09.2022
Project Title:	Alpha testing of the Jeeves Android app		
Researcher(s):	DanielRough, FergusNeville, Maximilian Stieler	Supervisor(s):	Aaron Quigley/David Harris-Birtill

Ethical amendment approval does not extend the originally granted approval period of three years, rather it validate the changes you have made to the originally approved ethical application. If you are unable to complete your research within the original five year validation period, you are required to write to your School Ethics Committee Convener to request a discretionary extension of no greater than 6 months or to re-apply if directed to do so, and you should inform your School Ethics Committee when your project reaches completion.

Any serious adverse events or significant change which occurs in connection with this study and/or which may alter its ethical consideration, must be reported immediately to the School Ethics Committee, and an Ethical Amendment Form submitted where appropriate.

Approval is given on the understanding that you adhere to the 'Guidelines for Ethical Research Practice' (http://www.st-andrews.ac.uk/media/UTRECguidelines%20Feb%2008.pdf).

Yours sincerely
Tristan Henderson

ethics-cs@st-andrews.ac.uk

The University of St Andrews is a charity registered in Scotland: No SC013532

APPENDIX C

APPENDIX-C STUDY MATERIALS

This Appendix contains written materials provided to participants in the usability studies conducted in Chapter 5, as well as interview guides for those conducted in Chapters 6 and 7. Study data, interview transcription, and video material is provided in the accompanying CD.

C.1 Usability Study 1

C.1.1 Pre-study Questionnaire

Participant ID:___

1. Age _____
2. Gender______
3. Spoken language: ______________________
4. Subject of study at university:______________________

5. How long do you use a computer for on an average week?
◯ Less than 10 hours
◯ 10-20 hours
◯ 20-30 hours
◯ More than 30 hours

6. How long do you spend playing video games on an average week?
____

7. Do you have any experience with computer programming? (Delete as appropriate): **YES/NO** (If **NO**, go to question 11)

8. Which programming languages do you have experience with?
____

9. Of these languages, which one do you consider yourself the most proficient in?
____

10. In terms of this language, how confident do you feel as a programmer?

Complete beginner	**Advanced beginner**	**Competent**	**Advanced**	**Expert**

11. Have you ever used a smartphone application that:
a) Sends you reminders of things to do? **YES/NO**
b) Collects and analyses your smartphone data? **YES/NO**
c) Automatically performs tasks, such as turning on WiFi or muting sounds? **YES/NO**

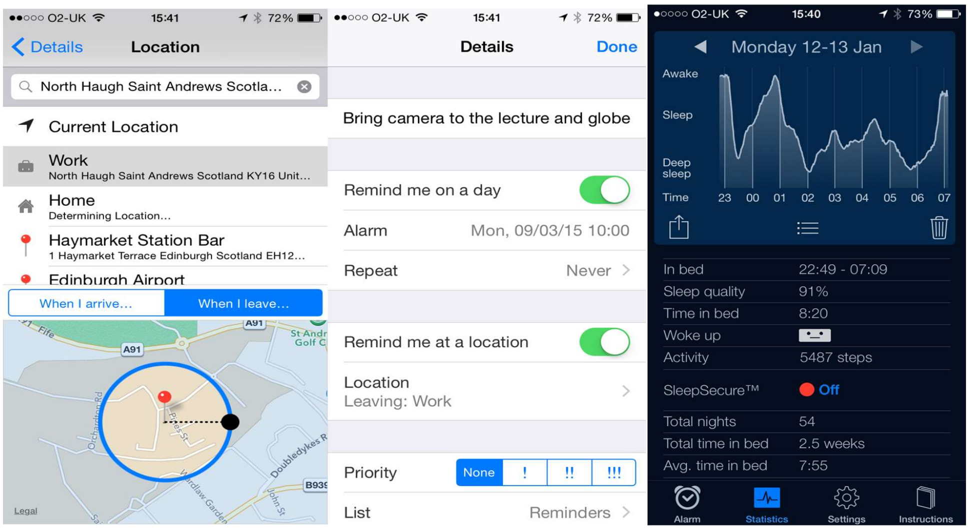
12. If the answer to any of the above was **YES**, please mention some relevant applications you have used.
____
____

13. Are you familiar with the “Experience Sampling Method” or “Ecological Momentary Assessment”? **YES/NO**

C.1.2 Jeeves Introduction

Introduction

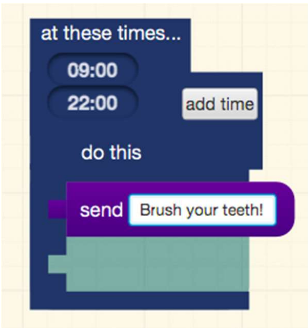
Many apps available to smartphone owners are becoming "context-aware" – meaning that they react differently based on their surroundings. These apps include those for sending reminders, automating actions, or monitoring aspects of your life. Example screenshots of such apps are shown below.



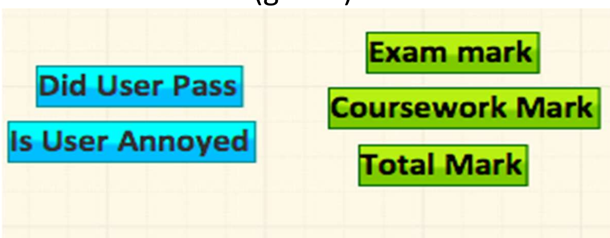
These apps work by waiting for certain **triggers**, then performing **actions** when these occur. Triggers are time-based or sensor-based.

The interface you will use today allows you to create basic trigger-action apps that can run on Android smartphones.

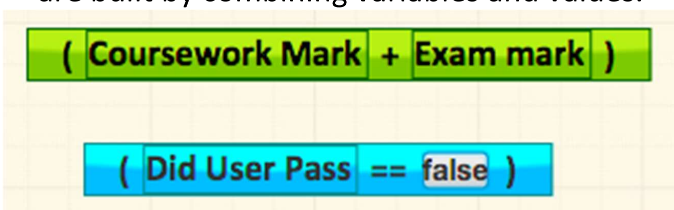
- Below are two simple examples. The **time trigger** fires at 9am and 10pm, the **action** reminds the user to brush their teeth.
- In the other example, the **sensor trigger** fires when the light sensor detects darkness, the **action** mutes the user's phone.



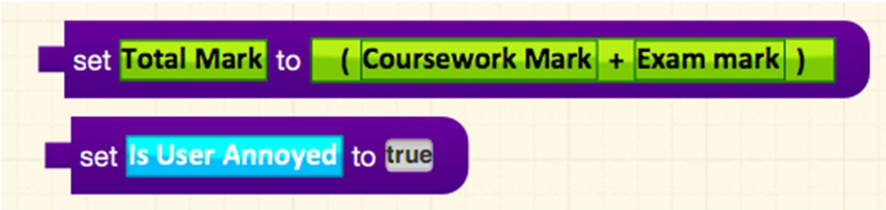
Variables are used in the interface to represent true/false values (blue), or numbers (green).



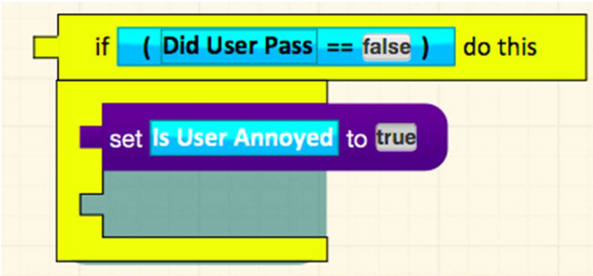
Expressions also represent true/false values, or numbers, and are built by combining variables and values.



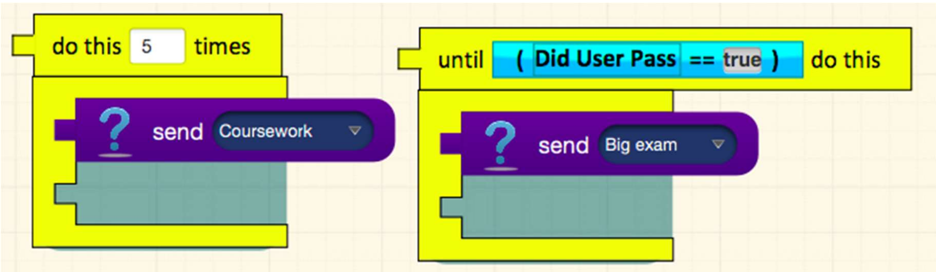
Update Variable actions can be combined with variables and expressions to set them to different values



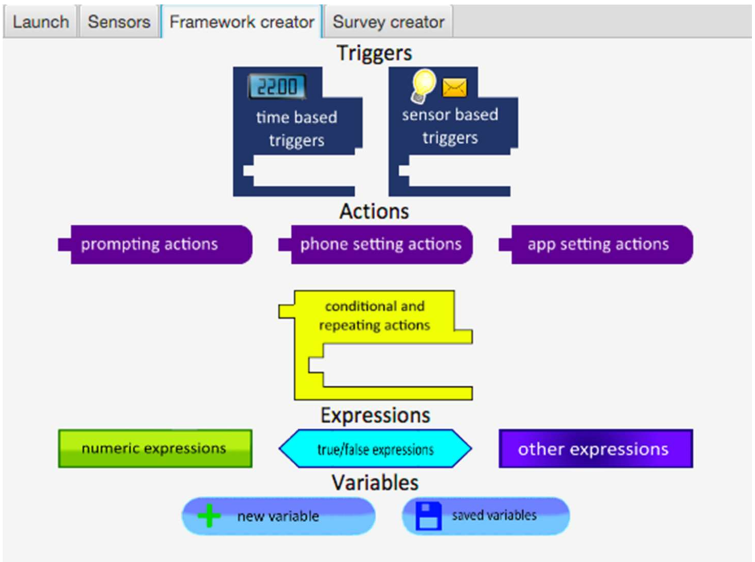
If conditions are special actions that only execute their containing actions if a condition is true.



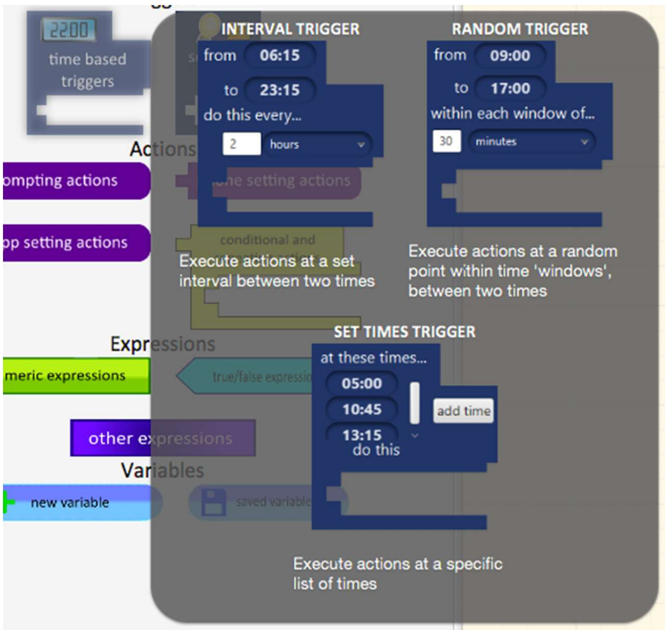
Do X times Loops and **Do Until Loops** are also special actions, that execute their containing actions either a set number of times, or until a condition is true.



All of the above components are accessible through the 'Framework creator' tab.

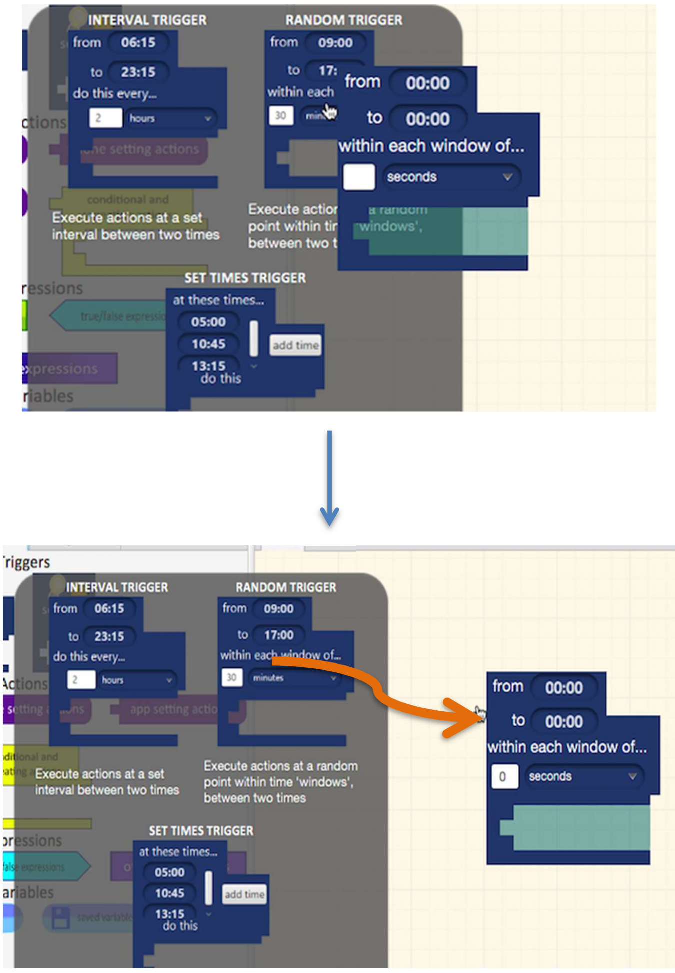


Each button on this tab represents a category of components that can be used. To make the specific components available, **click** on the button to bring up a menu like the one below.



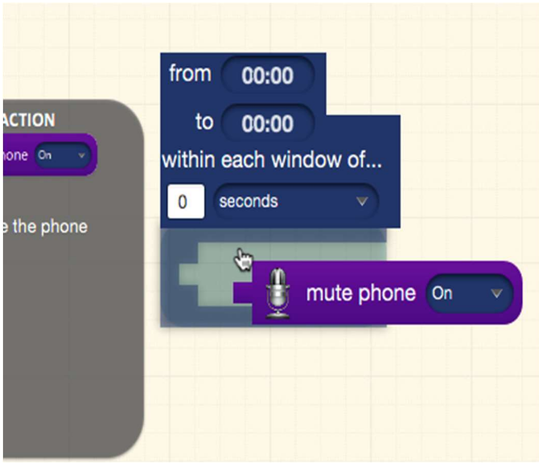
DRAGGING AND DROPPING COMPONENTS

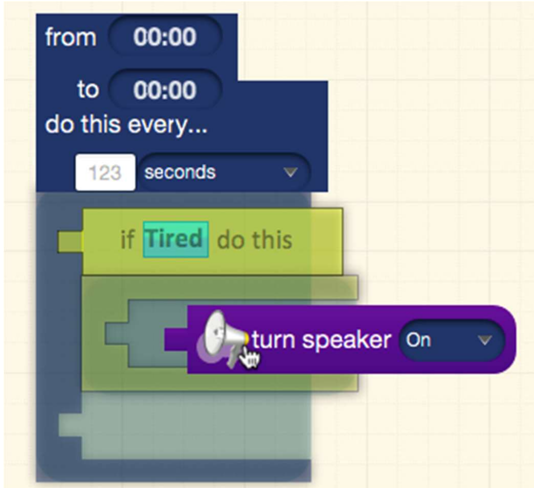
The components in the menu will show on hovering, and can be clicked, dragged and dropped onto the canvas like so.



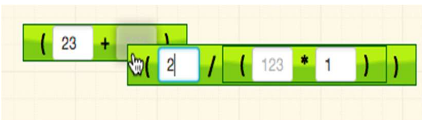
DROPPING COMPONENTS INSIDE EACH OTHER

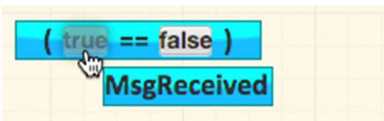
Actions can be dragged into the brackets of triggers or **conditional actions**.

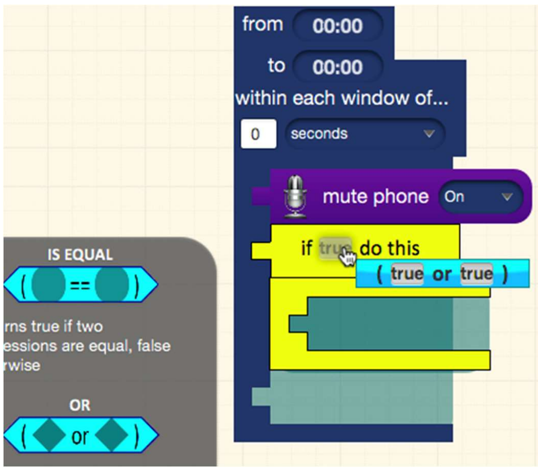




Expressions and **variables** can be dragged inside each other, or into **conditional actions**

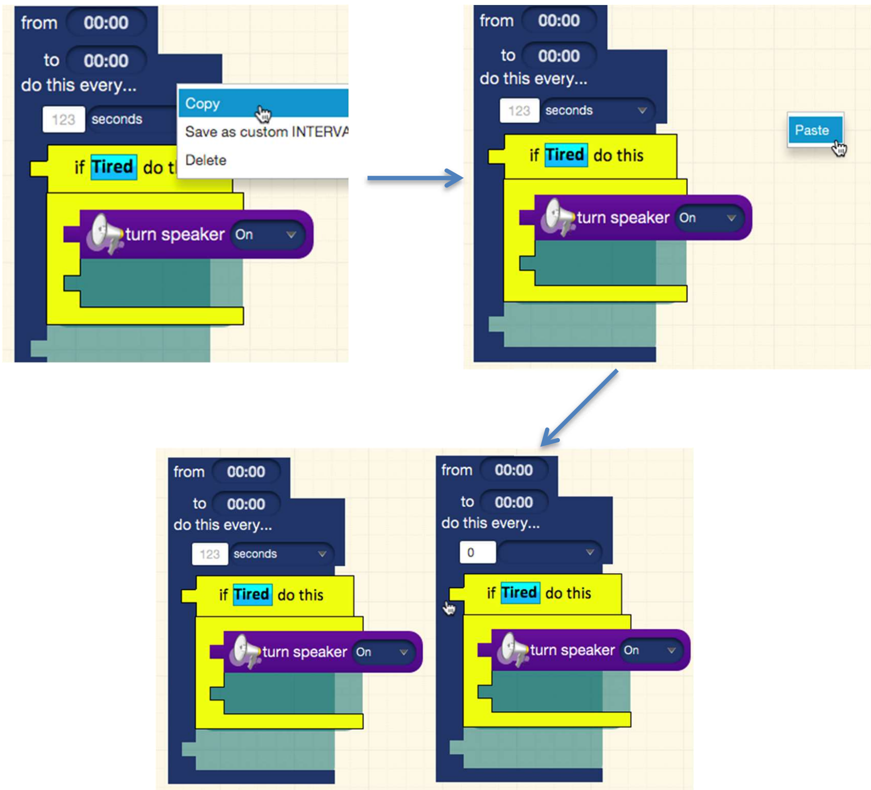






COPYING/DELETING COMPONENTS

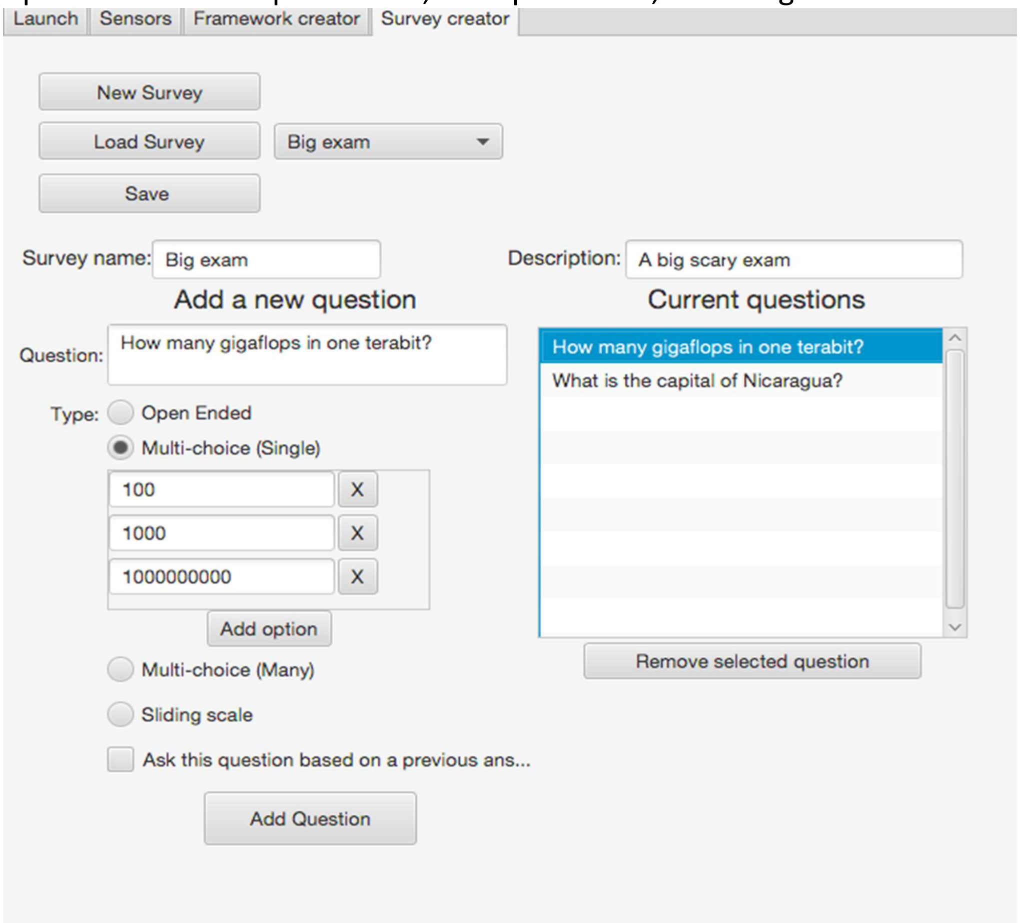
The menu for copying and deleting can be accessed by right-clicking on a component.
After copying, right-clicking on the canvas brings up the menu to paste a new copy.



Selecting 'Delete' from this menu removes the component from the canvas.

SURVEYS

Finally, **surveys** can be created in the "Survey creator" pane, which are sent to the user to collect their information. Survey questions can be open-ended, multiple choice, or sliding scale.

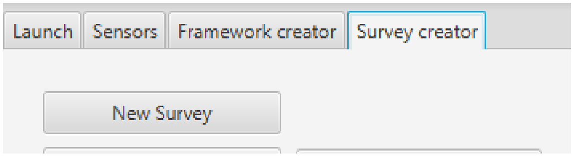


C.1.3 Study Tasks

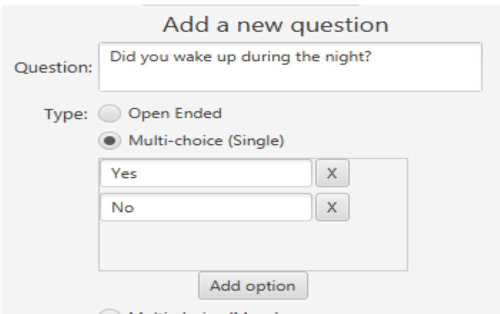
Scenario Overview

Your sleep has been poor recently and you'd like to find out why. You decide to design an app that lets you correlate your sleep with different events throughout your day.

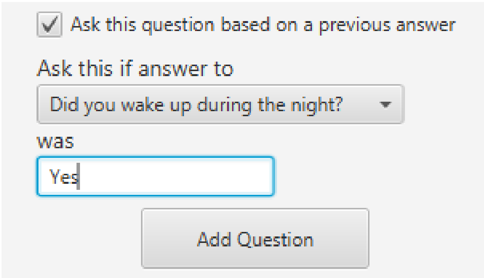
1. First, you'll need a questionnaire to answer when you wake up.
    - Create a new survey by selecting the "**Survey creator**" tab, then clicking the "**New Survey**" button.
    - Give your survey a relevant **Name** and **Description**



2. Add a new question with the following parameters:
    - Question text: "Did you wake up during the night?"
    - Question type: Multi-choice (Single)
    - Options: "Yes" and "No"
    - Then press the "**Add Question**" button to add this to your list of questions.

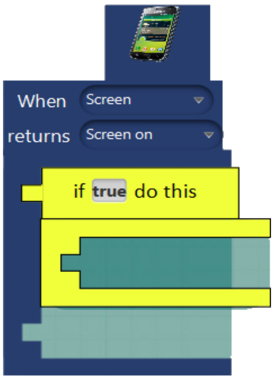


3. You would like a question that asks how many times you woke up, but *only if* your answer to the previous question was "Yes". To do this, add a question with the following parameters:
    - Question text: "If so, how many times?"
    - Question type: Open-ended
    - Then, check the box marked "**Ask this question based on a previous answer**".
    - In the "**Ask this if answer to**" dropdown box, find and select the text of your previous question, (i.e. "Did you wake up during the night?")
    - Type "Yes" into the text box, and add the question as before.

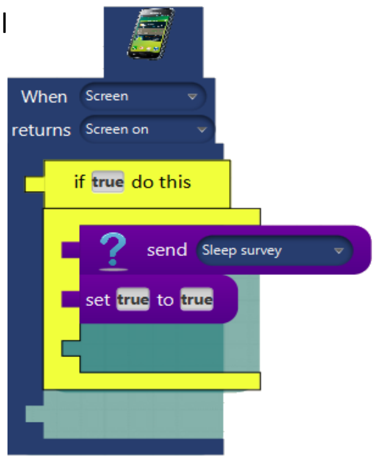


4. You're all done with the survey for now.
   - Save your survey by clicking the "**Save**" button. The form will clear, and your survey is now saved!
   - Return to the "**Framework Creator**" tab to start configuring the application.

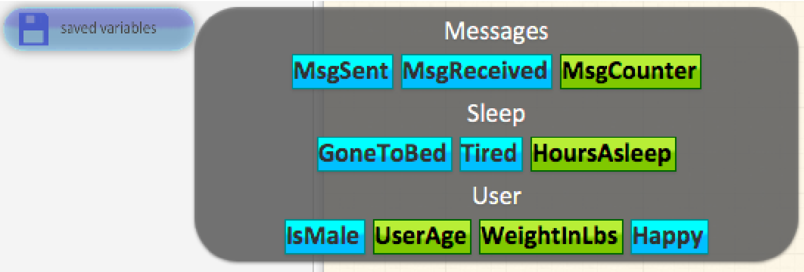
5. You'd like this survey to be triggered when you first look at your phone in the morning.
   - Drag a new **SOFTWARE SENSOR TRIGGER** onto the canvas, and set the dropdown options so that it fires when the screen is turned on.
   - Inside this trigger, you want a condition that checks whether you have gone to bed. Drag an **IF CONDITION** into this trigger from the **conditional and repeating actions** menu.



6. Inside the **IF CONDITION**, you'll need two actions that will be executed if the condition is true.
   - Find the **SURVEY ACTION** and add it to the if condition.
   - Then find the **UPDATE VARIABLE ACTION** and add it to the condition too.



7. Next, you need a variable to represent whether the user has gone to bed or not.

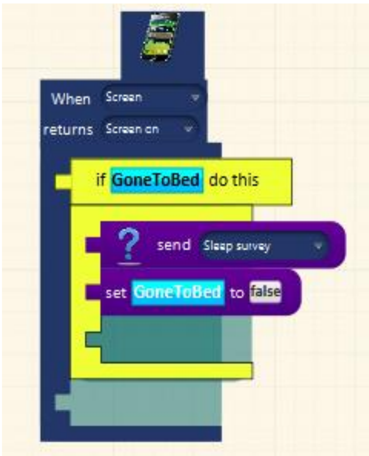


-Click on the "**saved variables**" button, and look for a blue variable called "*GoneToBed*". (**Blue** variables are **true/false** values. **Green** variables are **number** values).

-Drag this variable into the **if...do this** condition part of your **IF CONDITION**. This means that the actions within this **IF** are only executed if "*GoneToBed*" is true.

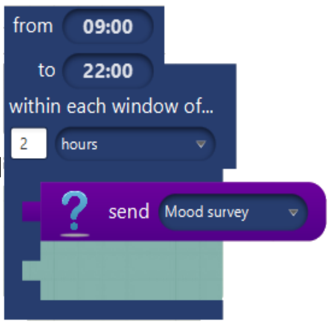
8. The last action of this trigger will set the "*GoneToBed*" variable to be false after the survey is sent. To do this:

- Drag another copy of "*GoneToBed*" from the **saved variables** section, and place this in the left side of your **UPDATE ACTION** as below.
- Click the 'true' button on the right hand side of the action, which sets it to 'false', and you should end up with something like the image on the right.



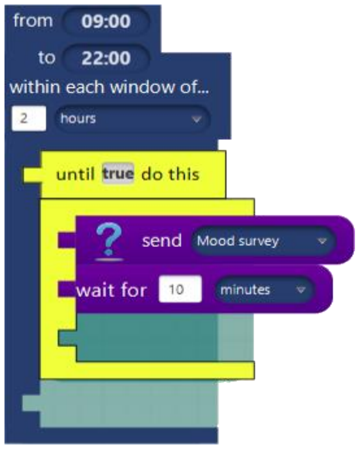
9. To check your mood through the day, you'll need a trigger that randomly fires every few hours.

- Find the **RANDOM TRIGGER** in the "**time triggers**" menu and add it to the canvas. Set the parameters so that it fires every 2 hours between 9am and 10pm.
- Add another **SURVEY ACTION** to this trigger, and select "*Mood survey*" from the dropdown.



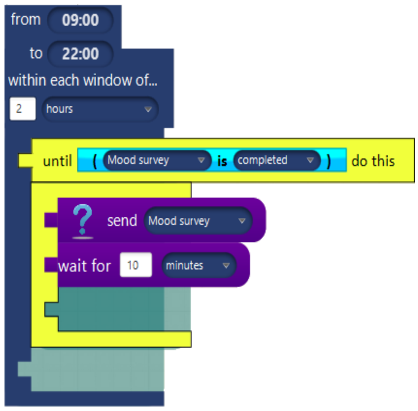
10. It's possible that you're unable to complete the survey at the time it is sent, but you want the survey to be sent until it is completed.

- Drag a **DO UNTIL** loop into the trigger from the "**conditional and repeating actions**" menu
- Drag your **SURVEY ACTION** so that it's now *inside* the loop.
- Inside the loop, also add a **WAIT ACTION** to say when to test again, and set the parameters to wait for 10 minutes.



11. At the moment this loop runs until "true". To set a condition:

- Find the **SENSOR EXPRESSION** from the "**other expressions**" menu and drag it onto the canvas.
- Set this expression to be "*Mood Survey is completed*" using the dropdown boxes.
- Drag the expression to the condition area of your loop as shown, and you're done!



Please now turn over the page and begin Part 2 of the study.

PART 2 – READING AND UPDATING TASKS

In the following part of the experiment, you are asked to read and briefly describe what a previously created application is designed to do as accurately as you can.
You will then be given the correct answer, and a task to update the application in some way.

1. Open the "**Part 2**" app, then answer the following question:

In a couple of sentences, describe what you think this app is designed to do.

____
____
____
____
____
____

Once you have answered, please ask the researcher for Part 2 of the study script.

Updating tasks

The application works well, but you also want to be able to handle what happens when a phone call is received while you're driving.

1. Add another **trigger** that detects when a **call has begun**. If the user is driving, turn on the **speaker phone.**

2. Create a **survey** with one question that asks how many messages the user plans to respond to. The question should have the options “ALL”, “SOME”, and “NONE”. Send this survey **AFTER** prompting that they have SMSs to reply to.

Please now begin Part 3 on the next page.


Ethel **Susan**

Use case: Ethel and Susan

Susan's mother Ethel lives on her own and has been finding it difficult to remember when to take her medication. She has to take it at three set times every day, and without someone to remind her, she often forgets.

Ethel is recovering from surgery, and Susan worries that she gets stuck in bed or the bath, and is unable to contact anyone. Recently she has been phoning her mum every few hours to check she’s okay.

Ethel has recently purchased a smartphone, and Susan would like an app that would help them both. Ethel wakes up at 8am and goes to bed at 8pm.

Design an app that reminds Ethel to take her medication in the morning, afternoon and evening. The app should also ask her every few hours if she's okay. If she doesn't respond after a certain time, the app should let Susan know that Ethel has been unable to respond.

C.1.4 Post-Study Survey

1. I think that I would like to use Jeeves frequently.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
2. I found Jeeves unnecessarily complex.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
3. I thought Jeeves was easy to use.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
4. I think that I would need the support of a technical person to be able to use Jeeves.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
5. I found the various functions in Jeeves were well integrated.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
6. I thought there was too much inconsistency in Jeeves.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
7. I would imagine that most people would learn to use Jeeves very quickly.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
8. I found Jeeves very cumbersome to use.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
9. I felt very confident using Jeeves.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
10. I needed to learn a lot of things before I could get going with Jeeves.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
11. I found it easy to create a survey	strongly disagree	◯	◯	◯	◯	◯	strongly agree
12. I found it easy to modify triggers, actions and expressions.	strongly disagree	◯	◯	◯	◯	◯	strongly agree
13. I like the look of the visual components	strongly disagree	◯	◯	◯	◯	◯	strongly agree
14. I think the trigger-action layout is easy to understand	strongly disagree	◯	◯	◯	◯	◯	strongly agree
15. I feel confident that I could create basic applications on my own	strongly disagree	◯	◯	◯	◯	◯	strongly agree

C.2 Usability Study 2

C.2.1 Pre-study Questionnaire

Participant ID:___

1. How long do you use a computer for on an average week?
    a. Less than 10 hours ☐
    b. 10-20 hours ☐
    c. 20-30 hours ☐
    d. More than 30 hours ☐

2. How long do you spend playing video games on an average week?
    ____

3. Do you have any experience with computer programming? (Delete as appropriate): **YES/NO** (If **NO**, go to Question 10)

4. Which programming languages do you have experience with?
    ____

5. Of these languages, which are you most comfortable with?
    ____

6. In terms of this language, how confident do you feel as a programmer?

Complete beginner	Advanced beginner	Competent	Advanced	Expert

7. Before being notified of this workshop, were you familiar with the "Experience Sampling Method" or "Ecological Momentary Assessment"? **YES/NO**

8. If **YES**, how have you seen it being used?
    ____
    ____

9. Age ___
10. Gender _________
11. Spoken languages ____

Jeeves - A Visual Programming Environment for Experience Sampling

1

Schedule

Now-9.30: Introduction to EMA and Jeeves

9.30-10.45: Jeeves installation, programming exercises 1-4

10.45-11.00: End-of-workshop surveys, open feedback

11.00-12.00: Lecture on health technology from Aaron

2

EMA: Why use it?

- Or “Ecological Momentary Assessment”
- “In-situ” data capture
- Minimise recall bias
- Long-term data capture

The old: Pen-and-paper diaries

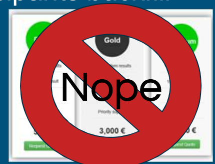
The new: Mobile EMA apps

3

EMA: Why *not* use it?

- Programming is *hard*
- Tools are *expensive*
- Pen and paper is *cumbersome…*
- …and participants *backfill*

✗ External developer?



✗ Commercial creation tool?

✗ Stick with paper?

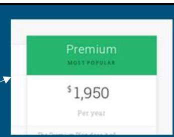
Who is Jeeves?

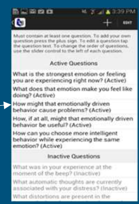
- Java
- End-User
- Environment for
- Visual
- Experience
- Sampling

5

Other options

- The expen$ive
  - movisensXS
  - MetricWire
  - LifeData
- The restrictive
  - iPromptU
  - PIEL
  - Paco
- The unapproachable
  - Ohmage
  - MyExperience
  - Momento

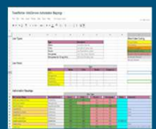




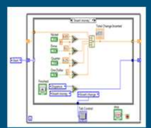
6

Visual Programming

Scratch



Spreadsheets



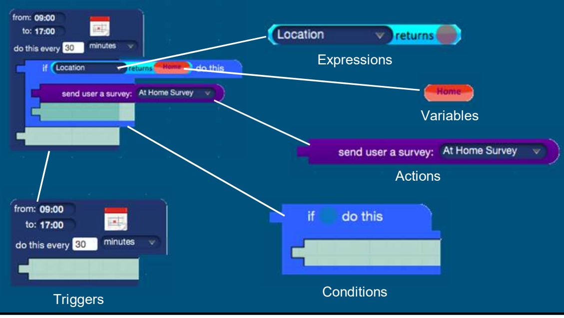
LabVIEW

App Inventor

7

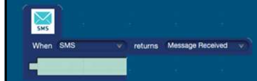
Our solution

Block-based + EMA = Jeeves

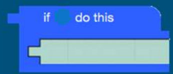


8

Drag-n'-drop bits



Triggers
- Event-based
- Time-based



Conditions
"If expression is true… do the following"

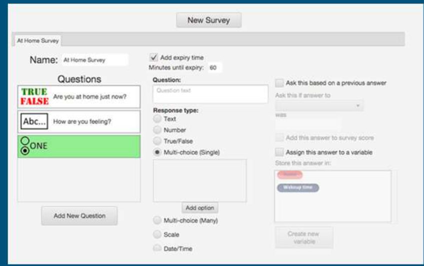
Variables- User-specific info
- True/False
- Numeric
- Time
- Date
- Location

Actions
- Prompting messages
- Sending surveys
- Adjusting settings

Expressions
Combining variables
- True/False

9

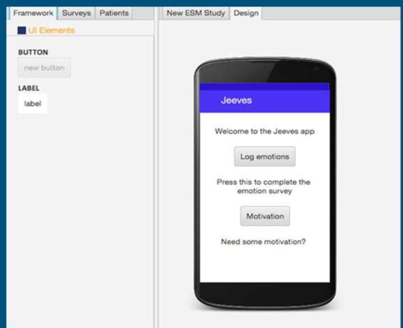
Other bits - Surveys



Key component of ESM
- Open questions
- Multi-choice questions
- Scale questions
- Conditional logic
- Branching

10

Other bits - UI Design



How does it look?
- Add labels
- Add buttons
- Trigger on button press

11

Thanks for listening!

Now
- Java Runtime installation (if you haven't got it yet)
- Jeeves download
- Step-by-step walkthrough

http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html

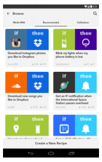
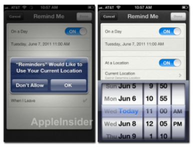
http://djr53.host.cs.st-andrews.ac.uk/download.zip

http://sachi.cs.st-andrews.ac.uk/walkthrough-part-1/

12

C.2.3 Jeeves Walkthrough

Many apps available to smartphone owners are becoming increasingly "context-aware" – meaning that they react differently based on their surroundings. These apps include those for sending reminders to oneself, automating simple actions, or monitoring different aspects of your physical/mental health.

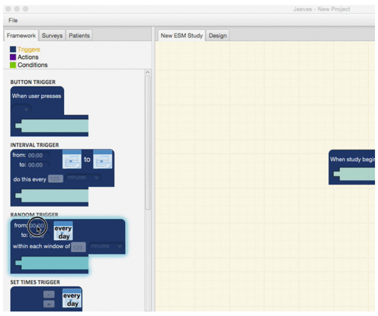
This also includes EMA apps, whereby participants are prompted to complete surveys based on time, or other external factors such as location or phone usage. Clinicians and researchers can benefit from data collected by such apps, but without any programming knowledge,it is difficult or expensive for them to do so.

With the Jeeves environment, we hope to give researchers full control of creating and deploying such apps for Android smartphones, with **no coding experience required**. The easiest way to get to grips with the environment is to walk step-by-step through the creation of a basic application. Take the following study description from a publication in the Journal of Psychiatric Research:

"The devices were preprogrammed to beep randomly 12 times/day (10 am/10 pm) to elicit 24 experience samples over a two-day period. Upon hearing the beeps, participants were instructed to complete a brief questionnaire, which also included four questions about current emotions – sadness, anger, anxiety, and happiness. For each question, participants were asked to rate their current experience using a graphical slider. Responses were represented in the output as a value between 1 ("not at all") and 100 ("very much")"

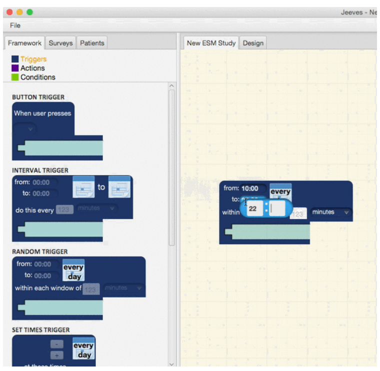
To help walk you through, we've included some animations. Some of these are a bit patchy and ugly but they should hopefully give you an idea of what's going on.

For this simple example, we'll need a RANDOM TRIGGER. Drag it onto the canvas like so:



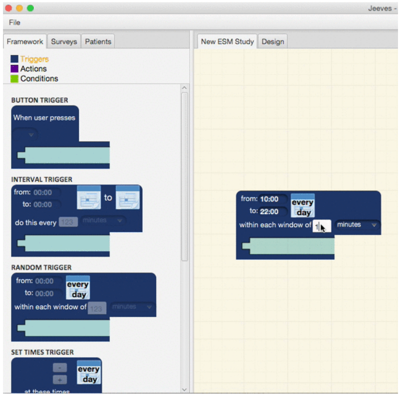
(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/randomtrig.gif)

We want it to beep within the hours of 10am and 10pm, which can be set by clicking on the times and editing them like so:



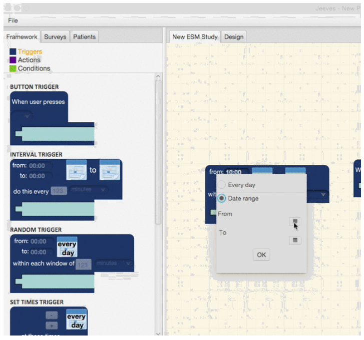
(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/changetimes.gif)

We want this trigger to fire roughly every hour, which we do by adjusting the trigger parameters:



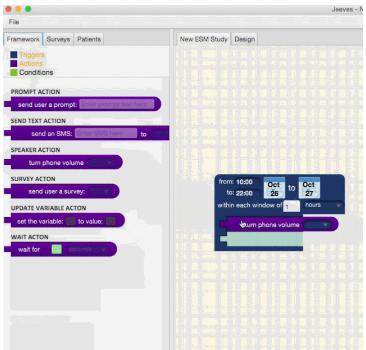
(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/changerate.gif)

The study only needs to run over a two day period, so we'll need to adjust the dates. Click on the calendar to make changes to the date range, and select two consecutive dates:



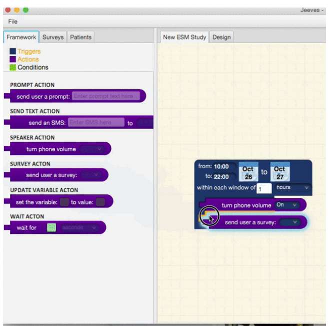
(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/changedates.gif)

When this trigger is 'fired', we want it to alert the user of a survey using an audible beep. First, we'll need to ensure that the user's sound is on, which we can do using the SPEAKER ACTION. Drag this into the trigger's brackets like this:



(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/addvolume.gif)

Next we'll send the user a survey, so grab the SURVEY ACTION and drag it into the trigger like this:



(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/addsurvey.gif)

Head to the next page to continue this walkthrough. (http://sachi.cs.st-andrews.ac.uk/research/interaction/visual-programming-languages/walkthrough-part-2/)

C.2.4 Study Tasks

In your download, you'll find the 'Studies' folder that contains five research publications that have used EMA to collect data from their participants. (Some of these studies will refer to 'ESM' but they mean the same thing.)

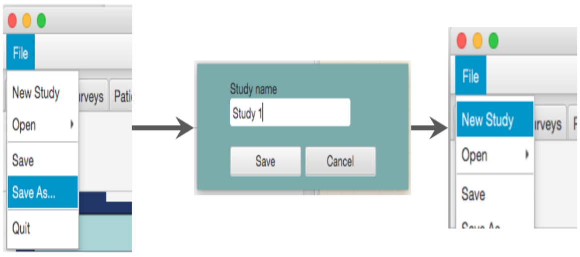
The aim of this exercise is to attempt to replicate the studies described in these publications.

(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/replication_publications1.jpg)

The necessary information has been highlighted in yellow in each of the publications so that you don't have to read through the whole thing! Information about the specification that can be ignored has been highlighted in blue.

The studies will not give full details on the survey questions asked. We ask you to use the details given, and your own imagination, to create suitable surveys.

When you have completed one, save it as 'Study X' where X is the number given in the filename, and move on to the next one.



(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/Screen-Shot-2016-10-09-at-01.00.10.png)

If you have any questions or are unsure how to proceed, please ask one of the researchers for assistance.

Once you're done, please move on to Exercise 3. (http://sachi.cs.st-andrews.ac.uk/exercise-3/)

In this exercise, we'd like you to compare Jeeves with the 'MovisensXS' environment.

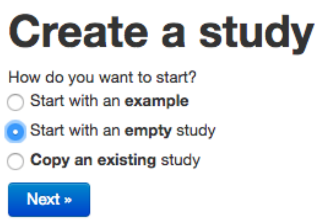
https://xs.movisens.com/login (https://xs.movisens.com/login)

To access this, the login credentials are as follows:

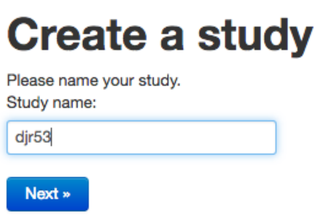
Username: djr53@st-andrews.ac.uk
Password: password

Once you've logged in, click the 'Add new Study' button.

Then, select the 'Start with an **empty** study' option



When asked to name your study, call it the ID that was given on your pre-workshop questionnaire.



The survey creation tool is accessed by clicking on 'Create forms', and the study configuration is accessed by clicking on 'Create sampling' as shown below.

(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/Screen-Shot-2016-10-10-at-17.59.02.png) (https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/Screen-Shot-2016-10-10-at-17.58.21.png)

Now try to create a study like the one described in the 'Study 2' paper in the previous exercise, which has been copied below:

"*Following the baseline assessments, adolescents were provided with smart phones that were programmed to beep three times a day for 30 consecutive days. Alarms were individually programmed to be compatible with each adolescent's normal waking hours as well as school schedules and other activities. The morning survey was scheduled between the times of 7 a.m. and 10 a.m., and took approximately 2.3 min to complete. The afternoon survey was scheduled between the hours of 2 p.m. and 5 p.m., and took on average 3.8min to complete. Finally, the p.m. survey was scheduled between the hours of 5 p.m. and midnight, and took on average 8.3 min to complete*"

Take your time and ask me if you need any help. Once you've completed this, save your study by clicking the icon in the top right corner, and click here to continue to Exercise 4. (http://sachi.cs.st-andrews.ac.uk/exercise-4/)

Now that we hope you are familiar with the Jeeves environment, and with the type of EMA studies that can be created with it, we would like you to spend 20 minutes designing and implementing **your own idea** from scratch.

Think about the research topic – what kind of problem would you like to solve? What kind of data would be valuable when collected 'in the moment' that would give more insight into this problem?

It's likely that you'll have ideas that can't be implemented with the current version of Jeeves – that's great! If you think of features you'd like included, please mention them in the feedback form that we'll hand out at the end of the workshop.

Create a new study for implementing your idea, and when you're done, save it as '**Exercise 4**'.

Again, if you have any questions, please ask one of the researchers for assistance.

(https://sachi.cs.st-andrews.ac.uk/wp-content/uploads/2016/10/leadImage.jpeg)

After this, you're all done! Once you've finished, please click the following links and complete both surveys. Thank you very much for participating.

Feedback 1: https://www.surveymonkey.co.uk/r/SM659F9 (https://www.surveymonkey.co.uk/r/SM659F9)

Feedback 2: https://www.surveymonkey.co.uk/r/SFMLQXS (https://www.surveymonkey.co.uk/r/SFMLQXS)

C.2.5 Post-Study Surveys

1. I think that I would like to use this system frequently.

strongly disagree	2	3	4	strongly agree

2. I found the system unnecessarily complex.

strongly disagree	2	3	4	strongly agree

3. I thought Jeeves was easy to use.

strongly disagree	2	3	4	strongly agree

4. I think that I would need the support of a technical person to be able to use Jeeves.

strongly disagree	2	3	4	strongly agree

5. I found the various functions in Jeeves were well integrated.

strongly disagree	2	3	4	strongly agree

6. I thought there was too much inconsistency in Jeeves.

strongly disagree	2	3	4	strongly agree

7. I would imagine that most people would learn to use Jeeves very quickly.

strongly disagree	2	3	4	strongly agree

8. I found Jeeves very cumbersome to use.

strongly disagree	2	3	4	strongly agree

9. I felt very confident using Jeeves.

strongly disagree	2	3	4	strongly agree

10. I needed to learn a lot of things before I could get going with Jeeves.

strongly disagree	2	3	4	strongly agree

1. I found the Jeeves survey creation pane easy to understand

strongly disagree	2	3	4	strongly agree
☆	☆	☆	☆	☆

2. I dislike the appearance of the visual components

strongly disagree	2	3	4	strongly agree
☆	☆	☆	☆	☆

3. I think the block-based layout is intuitive

strongly disagree	2	3	4	strongly agree
☆	☆	☆	☆	☆

4. I don't think I could create basic applications on my own

strongly disagree	2	3	4	strongly agree
☆	☆	☆	☆	☆

5. In Exercise 3, when you used movisensXS, how did you find this compared to Jeeves? Are there features you liked or disliked about it? Which would you rather use? Please comment freely below.

6. Finally, we'd welcome any other feedback on Jeeves or the workshop structure in general. If you have any thoughts, please write them in the comment box below.

7. Please enter your Participant ID here.

Done

C.3 Usability Study 3

C.3.1 Pre-study Questionnaire

Participant ID:___

1. Age _____
2. Gender______
3. Spoken language: ______________________

4. How long do you use a computer for on an average week?
◯ Less than 10 hours
◯ 10-20 hours
◯ 20-30 hours
◯ More than 30 hours

5. Do you have any experience with computer programming? (Delete as appropriate): **YES/NO** (If **NO**, go to question 9)

6. Which programming languages do you have experience with?
____

7. Of these languages, which one do you consider yourself the most proficient in?
____

8. In terms of this language, how confident do you feel as a programmer?

Complete beginner	Advanced beginner	Competent	Advanced	Expert

9. Are you familiar with the "Experience Sampling Method (ESM)" or "Ecological Momentary Assessment (EMA)"?
◯ **I have never heard of ESM/EMA**
◯ **I have heard of ESM/EMA but don't know what it involves**
◯ **I am familiar with ESM/EMA but have never used it in a user study**
◯ **I have used ESM/EMA in a user study before**

C.3.2 Study Tasks

Overview

You are going to make a new app that monitors a patient's stress level through the day. The app should periodically ask patients how they're feeling, and also allow them to log stressful events themselves.

Task 1 – Patient attributes

1. It is important to ensure that a patient is not woken up by notifications. At the very start of the study, get the patient's waking and sleeping times.

2. For two weeks, three times a day while the patient is awake, remind the patient to "Take time to practise your mindfulness exercises"

3. A patient's home environment could be an interesting context to sample stress. Tomorrow morning, ask the patient where they live.

4. It would be useful to know who else lives at home with the patient. Sample nearby Bluetooth devices when the patient returns home.

5. Save your study as "**My Study**" before moving onto Task 2.

Task 2 – Patient-initiated actions

As part of our study, we want patients to be able to log when they're at a location of interest, and to let the researcher know how they feel at this location.

1. Update your study so that the patient can log an interesting location.

2. After they log their location, the patient should be asked how stressed they feel, from 1 (not very stressed) to 5 (extremely stressed).

3. Save your study and move onto Task 3.

Task 3 - Compliance

Sometimes patients will fail to complete surveys. This can be a problem with a particular patient, or with all patients if the study is not setup correctly. We should be able to monitor and react to when this is happening.

1. Open the study "**TestStudy**" and check whether patients have been completing the surveys regularly.

.

2. If any patient is having a problem with compliance, ask them if they are having issues with the study.

3. All patients seem to have missed a few surveys. Check whether surveys are being sent too often or not enough.

4. Update the study to make sure that surveys are being sent at appropriate times.

5. Move onto Task 4.

Task 4 – Do Not Disturb Condition

Sometimes, even if the patient is awake, it might be a very bad time for them to receive a notification! They should be able to prevent this happening with a button that enables or disables 'Do Not Disturb' mode.

1. Open your study "**My Study**" that you begun in Tasks 1 & 2.

2. Jeeves should be able to remember that the patient is in Do Not Disturb Mode when a 'Do Not Disturb' button is pressed.

3. Patients should **not** receive their mindfulness reminder when they are in Do Not Disturb mode.

C.3.3 Post-study Questions

1. First of all, I'd like you to describe two of the most obvious issues that you recall having with Jeeves.

2. I'd also like you to describe two things you recall liking about Jeeves.

3. (If EMA familiar) You mentioned that you have familiarity with EMA. Could you go into more detail on that?

4. Was the user interface designer missing anything that you could imagine might be useful? What was it missing?

5. How did you find the idea of patient attributes? How would you change them to make them more understandable?

6. What is your first impression of the Patient compliance pane? Do you think this would be useful to a researcher?

7. Would you consider using Jeeves for collecting user data?

C.4 Interviews

C.4.1 Clinician Interview Script

Introduction

The purpose of this interview is to acquire feedback from potential end-users on the usefulness of my prototype platform for designing mobile apps for remote patient assessment. I want to try and get a better understanding of current working practices in the assessment and treatment of patients managing chronic conditions, any disadvantages to the current approach, and the possible supporting role of technology.

Later on I'd like to get your initial impressions of my developed platform as a means of supporting clinical practice.

Now, do you have any questions you'd like to ask before we begin properly?

Current Practices (15 minutes)

I would like to start off by asking about your current practices, specifically with regard to patients with ongoing conditions. Could you describe the structure of an appointment with a patient managing a chronic problem?

1. How does the time of the appointment affect the assessment and treatment?
2. What level of detail do **patients** give **you** in their appointments?
3. How regularly do you see such patients?
4. What kind of resources do you provide to patients during their appointments?
5. How are patients encouraged to manage their condition outside the clinic?
6. Where would you like to improve on the current process?

Technology Familiarity (10 minutes)

Moving on from current practices, I'd now like to talk about the future, and specifically about technology's role in improving chronic healthcare. What experience do you have with health technology?

1. In your experience with patients (managing chronic conditions), what would be a useful application of mobile health technology?
2. What do you think patients would want from a mobile application for monitoring their health?
3. What barriers would you face in beginning to use a mobile health application in your clinical practice?
4. How do you feel these barriers could be overcome?

Impressions of Jeeves and its real-world utility - 25 minutes

I'm going to show a couple of brief videos that demonstrate the functionality of Jeeves, and then after that I'd like us to discuss whether Jeeves could be useful in clinical practice, and how. I'd like to emphasise that you won't be hurting my feelings in any way. All feedback, positive or negative, is very useful

Follow-up questions:

- What is your general impression of the platform?
- Where do you think Jeeves could fit into clinical practice?
- Who do you think Jeeves would be most useful for?
- What barriers do you perceive to clinicians using Jeeves?
- How could Jeeves be improved?
- What do you think patients would need to maintain use of the Jeeves app?
- Which messages would you send to patients?

Possible probes:

- Detecting early signs
- Assisting in self-monitoring
- Patient independence
- Understanding fluctuations
- Improving communication during and between appointments
- Provision of information
- Informing shared decision-making

Wrap-up and Conclusion - 5 minutes

Summarising all the things we've discussed today, what would you say is the most important factor for you in beginning to use a platform like this?
Finally, is there anything you feel has not been addressed in our discussion today?

C.4.2 Psychologist Interview Script

Introduction

The purpose of this interview is to acquire feedback from potential end-users on the usefulness of my prototype platform for designing mobile apps for experience sampling. In summary, I want to try and get a better understanding of how you currently conduct research, what barriers you face in your research methods, and your initial impressions of the Jeeves platform as a means of supporting your research.

Now, do you have any questions you'd like to ask before we begin properly?

Current Practices (15 minutes)

First, I want to learn more about your current research practices, to contextualise the rest of our interview.

1. As I understand from recent work of yours I looked at, your research is focused on ***. Can you describe how you collect data from participants in this context?

2. Could you describe your reasons for employing these research methods? What alternatives were available? What were the positive aspects of employing this method?

3. What are the difficulties you've encountered with this kind of research?
    - Where were participants? When was the data collected? How was it collected?

4. What experience do you have with ESM?
    - (if used) Can you explain the study you conducted in more detail? What difficulties (if any) did you encounter in employing ESM?
    - (if not used) What would motivate you to use ESM in your research? What difficulties could you foresee in using ESM with your participants?

Technology Familiarity (10 minutes)

I would now like to move onto your experience and familiarity with software and technology within your research, and how it fits into your current practice.

5. What kind of software do you use in your research practice?
    - Could you describe your experience with learning and using this software?
    - What is the learning curve of this software?
    - Are there any particular likes or dislikes you have about the user interface?

6. What influences your motivation to use a software application in your research?

7. Are you familiar with any technologies available for conducting experience sampling research? If so, which ones?

- Have you considered using these...why/why not? If not, how do you envision such software would work?
- Considering your previous research, what requirements would this software need?

Use of Jeeves (30 minutes)

For the remainder of our time, I'd like to show you the Jeeves platform, and get your feedback on its usage. I'm going to walk you through a couple of examples that demonstrate the functionality of Jeeves, and then after that I'd like us to discuss whether Jeeves could be useful to you, and how. I'd like to emphasise that you won't be hurting my feelings in any way. All feedback, positive or negative, is very useful to me. If you have any questions as I demonstrate Jeeves, please feel free to ask.

8. What is your general impression of the Jeeves platform?
9. Where do you think Jeeves could fit into your work?
10. Who do you think Jeeves would be most useful for?
11. What barriers do you perceive to psychology researchers using Jeeves?
12. How could Jeeves be improved?
13. What do you think participants would need to maintain use of the Jeeves app?

Wrapping up (5 minutes)

Given our time, I'd like to begin wrapping up the interview now. I'd like to tie things together with a couple of questions, and then I'd be glad to answer any further questions

14. In summary, what are the main decisive factors towards whether you would use Jeeves in your research practice?

15. I would like to follow up with you on this. Would you be interested in taking part in a formal usability study over the next couple of weeks?

16. Thank you very much for your time. Do you have any other questions for me?

C.4.3 Post-Case Study Interview Script

Usefulness

1. First of all, do you have any general comments on your experience with running this study?
2. How did using Jeeves affect the overall design and data collection *compared to your previous studies*?
3. What do you feel was the most challenging part of running this study?
4. Has this influenced your intention to conduct more experience sampling research in the future?
5. If I had not introduced you to Jeeves and you had not used it before, what would have been the main factors you would have considered in deciding whether or not to use it?
6. Having used it and experienced it, what would influence whether or not you continue to use Jeeves in future?

Usability of blocks

1. How did you find the blocks-based programming interface? Can you make any general comments?
2. How does using Jeeves compare with other technology you use, Qualtrics for example?
3. What do you think of the length of time it took you to get used to Jeeves?
4. If we load up your study, I notice you've done this in a very top-to-bottom way. What influenced your decision to structure it like this?

Collaboration and support

1. When you've previously collaborated on studies, how do you tend to communicate? How did you find that compared to this particular study?
2. How would you like to be able to collaborate with other researchers using Jeeves?
3. How do you feel my involvement in the design and running process affected your use of Jeeves? If you were to run another study, what kind of support would you look for?
4. There are some features that you haven't used in your app. For example, attributes and if conditions. What would influence you to use these?
5. How important is it for you to be able to suggest new features to be implemented?

Participants

1. What do you think the benefits for participants were using Jeeves over your usual data collection approach?
2. What do you think the drawbacks for participants were, or would be in future?
3. Apart from their answers to surveys, is there any other information that you'd like to automatically collect on participants?
4. How important is it for you to be able to monitor participant compliance?

APPENDIX D

APPENDIX-D PREVIOUS JEEVES SCREENSHOTS

This appendix provides some screenshots from the two previous versions of Jeeves used in the first and second usability studies respectively. Screenshots of the most recent version of Jeeves are shown throughout Chapter 5, so are not included here. The purpose of these screenshots is to illustrate how Jeeves has evolved in response both to direct participant feedback, and the identified need for additional features through analysis of ESM literature.

The first two pages (Figures D.1- D.5) illustrate the first iteration of Jeeves prior to the first usability study, described in Section 6.2. The third page (Figures D.6- D.8) illustrate the second iteration prior to the second usability study, described in Section 6.3.

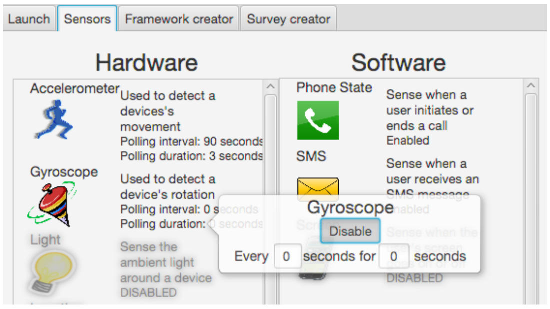


Figure D.1
Jeeves Sensor pane for selecting the frequency and granularity of sensing

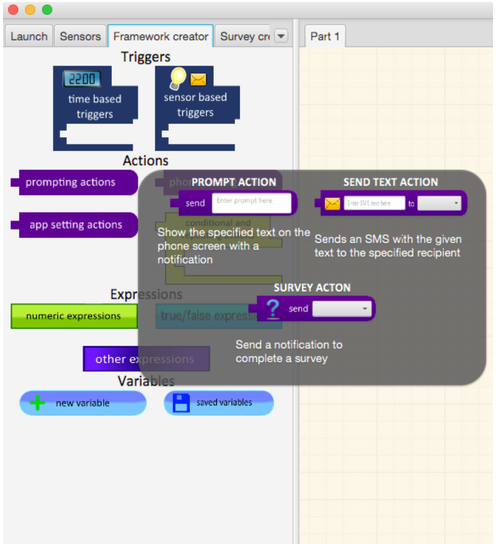


Figure D.2
Original "Blocks" view of Jeeves with blocks grouped into categories visible on hovering

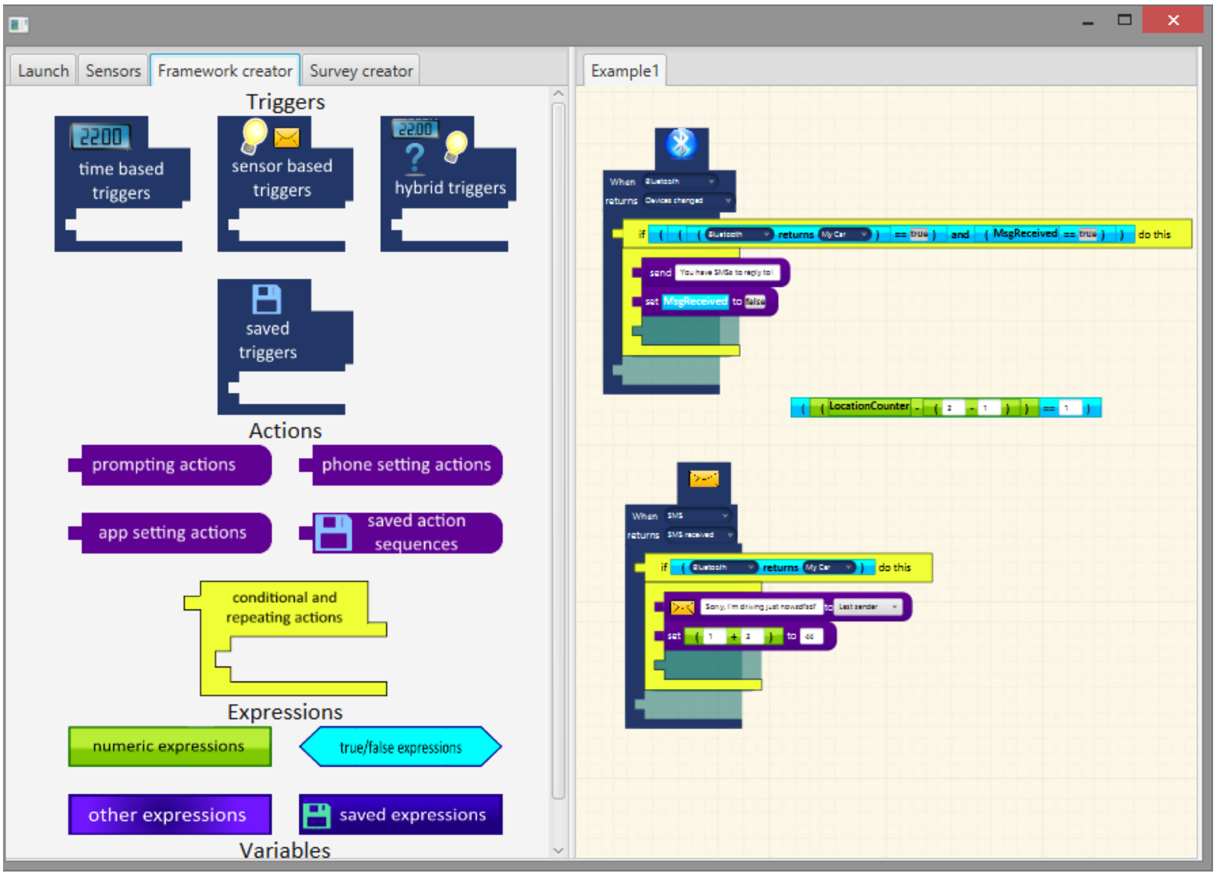


Figure D.3
Overall view of Jeeves prior to Study #1, showing categories of blocks, and an example specification

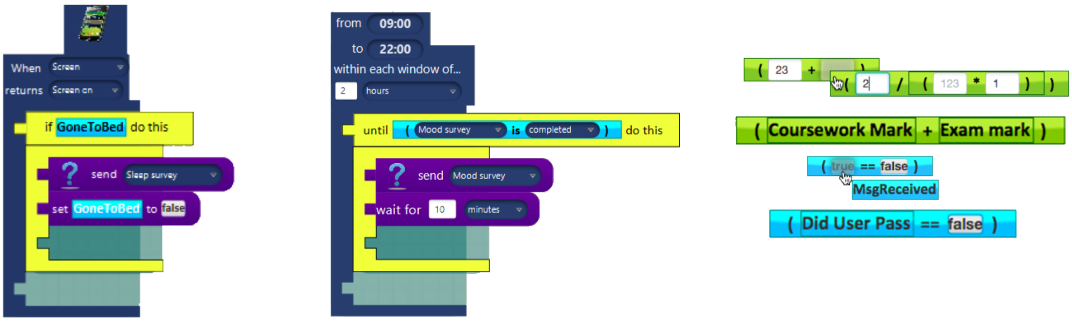


Figure D.4
Left and centre: example block combinations of triggers, conditions, actions and expressions. Right: complex expressions and depiction of their combination

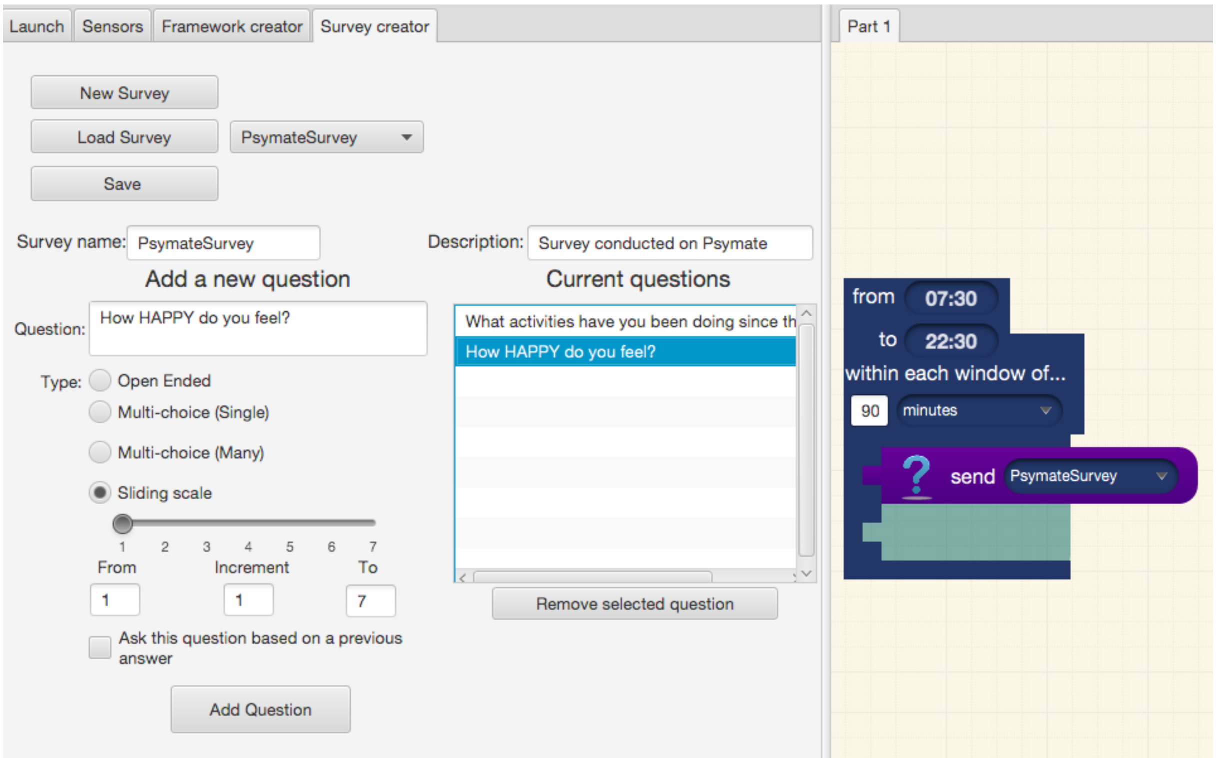


Figure D.5
Survey design view in the first iteration of Jeeves

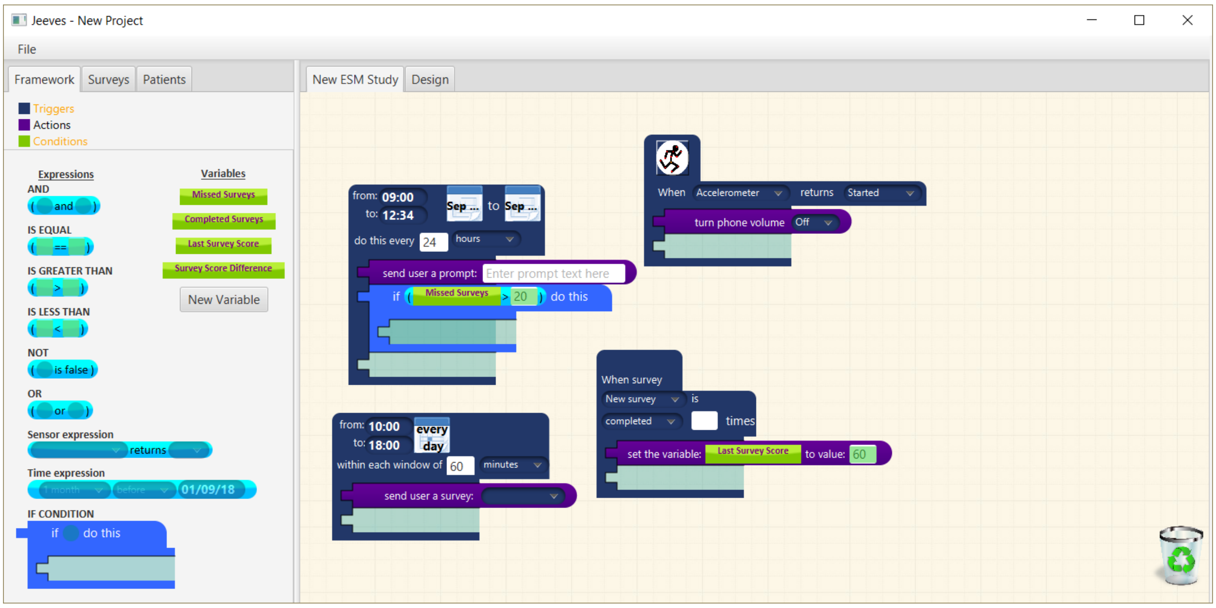


Figure D.6
Overall view of Jeeves prior to Study #2, showing categories of blocks, and an example specification

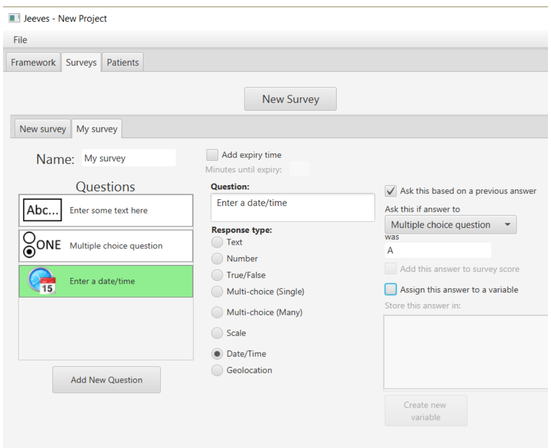


Figure D.7
Survey design view in the second iteration of Jeeves

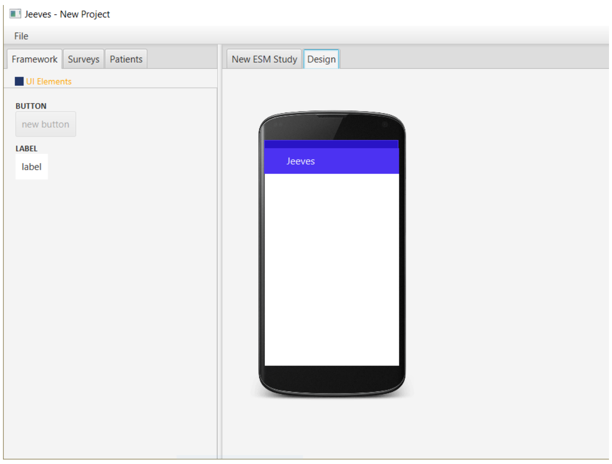


Figure D.8
Log Design view in the second iteration of Jeeves

References

[1] M. Weiser, "The computer for the 21st century," *Scientific American*, vol. 265, no. 3, pp. 94–105, 1991.

[2] C. Fullwood, S. Quinn, L. K. Kaye, and C. Redding, "My virtual friend: a qualitative analysis of the attitudes and experiences of Smartphone users: Implications for Smartphone attachment," *Computers in Human Behavior*, vol. 75, pp. 347–355, 2017.

[3] M. Kay, J. Santos, and M. Takane, "mHealth: New horizons for health through mobile technologies," *World Health Organization*, vol. 64, no. 7, pp. 66–71, 2011.

[4] Research2Guidance, "mHealth economics 2017 - current status and future trends in mobile health," 2017. [Online]. Available: https://research2guidance.com/product/mhealth-economics-2017-current-status-and-future-trends-in-mobile-health

[5] M. N. Kamel Boulos, A. C. Brewer, C. Karimkhani, D. B. Buller, and R. P. Dellavalle, "Mobile medical and health apps: state of the art, concerns, regulatory control and certification," *Online Journal of Public Health Informatics*, vol. 5, no. 3, p. e229, 2014.

[6] G. Miller, "The Smartphone Psychology Manifesto," *Perspectives on Psychological Science*, vol. 7, no. 3, pp. 221–237, May 2012.

[7] Z. W. Adams, E. A. McClure, K. M. Gray, C. K. Danielson, F. A. Treiber, and K. J. Ruggiero, "Mobile devices for the remote acquisition of physiological and behavioral biomarkers in psychiatric clinical research," *Journal of Psychiatric Research*, vol. 85, pp. 1–14, 2017.

[8] T. C. Christensen, L. F. Barrett, E. Bliss-Moreau, K. Lebo, and C. Kaschub, "A practical guide to experience-sampling procedures," *Journal of Happiness Studies*, vol. 4, no. 1, pp. 53–78, 2003.

[9] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the Experience-Sampling Method." *The Journal of nervous and mental disease*, vol. 175, no. 9, pp. 526–36, Sep 1987.

[10] A. Stone and S. Shiffman, “Ecological momentary assessment (EMA) in behavioral medicine,” *Annals of Behavioral Medicine*, vol. 16, no. 3, pp. 199–202, Jan 1994.

[11] J. Fahrenberg, “Ambulatory assessment: Issues and perspectives.” in *Ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies.* Ashland, OH, USA: Hogrefe & Huber Publishers, 1996, pp. 3–20.

[12] C. Napa Scollon, C.-K. Prieto, and E. Diener, “Experience sampling: Promises and pitfalls, strength and weaknesses,” in *Assessing Well-Being: The Collected Works of Ed Diener*, E. Diener, Ed. Dordrecht: Springer Netherlands, 2009, pp. 157–180.

[13] S. Shiffman, A. A. Stone, and M. R. Hufford, “Ecological momentary assessment,” *Annual Review of Clinical Psychology*, vol. 4, no. 1, pp. 1–32, 2008.

[14] T. J. Trull and U. Ebner-Priemer, “Ambulatory assessment,” *Annual Review of Clinical Psychology*, vol. 9, no. 1, pp. 151–176, 2013.

[15] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature*, vol. 533, no. 7604, pp. 452–454, 2016.

[16] H. Lieberman, F. Paternò, M. Klann, and V. Wulf, “End-user development: An emerging paradigm,” in *End User Development*, H. Lieberman, F. Paternò, and V. Wulf, Eds. Dordrecht: Springer Netherlands, 2006, pp. 1–8.

[17] J. Kjeldskov and C. Graham, “A review of mobile HCI research methods,” in *Human-Computer Interaction with Mobile Devices and Services*, L. Chittaro, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 317–335.

[18] D. Tetteroo and P. Markopoulos, “A review of research methods in end user development,” in *End-User Development*, P. Díaz, V. Pipek, C. Ardito, C. Jensen, I. Aedo, and A. Boden, Eds. Cham: Springer International Publishing, 2015, pp. 58–75.

[19] M. Csikszentmihalyi and R. Larson, “Validity and reliability of the experience-sampling method,” in *Flow and the Foundations of Positive Psychology*. Dordrecht: Springer Netherlands, 2014, pp. 35–54.

[20] S. Shiffman, “Ecological momentary assessment (EMA) in studies of substance use,” *Psychological Assessment*, vol. 21, no. 4, pp. 486–497, 2009.

[21] N. Bolger, A. Davis, and E. Rafaeli, “Diary methods: capturing life as it is lived,” *Annual Review of Psychology*, vol. 54, no. 1, pp. 579–616, 2003.

[22] M. B. Solhan, T. J. Trull, S. Jahng, and P. K. Wood, "Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall," *Psychological Assessment*, vol. 21, no. 3, pp. 425–436, Sep 2009.

[23] D. Ben-Zeev, G. J. McHugo, H. Xie, K. Dobbins, and M. A. Young, "Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group," *Schizophrenia Bulletin*, vol. 38, no. 3, pp. 396–404, 2012.

[24] I. Myin-Germeys, F. Peeters, R. Havermans, N. A. Nicolson, M. W. DeVries, P. Delespaul, and J. Van Os, "Emotional reactivity to daily life stress in psychosis and affective disorder: an experience sampling study," *Acta Psychiatrica Scandinavica*, vol. 107, no. 2, pp. 124–131, Feb 2003.

[25] T. J. Trull, M. B. Solhan, S. L. Tragesser, S. Jahng, P. K. Wood, T. M. Piasecki, and D. Watson, "Affective instability: measuring a core feature of borderline personality disorder with ecological momentary assessment," *Journal of Abnormal Psychology*, vol. 117, no. 3, p. 647, 2008.

[26] U. W. Ebner-Priemer and T. J. Trull, "Ecological momentary assessment of mood disorders and mood dysregulation," *Psychological Assessment*, vol. 21, no. 4, pp. 463–475, Dec 2009.

[27] A. A. Haedt-Matt and P. K. Keel, "Revisiting the affect regulation model of binge eating: a meta-analysis of studies using ecological momentary assessment," *Psychological Bulletin*, vol. 137, no. 4, pp. 660–681, 2011.

[28] J. T. Mitchell, M. F. Dennis, J. S. English, P. A. Dennis, A. Brightwood, J. C. Beckham, and S. H. Kollins, "Ecological momentary assessment of antecedents and consequences of smoking in adults with attention-deficit/hyperactivity disorder," *Substance Use & Misuse*, vol. 49, no. 11, pp. 1446–1456, 2014.

[29] G. F. Dunton, E. Dzubur, and S. Intille, "Feasibility and performance test of a real-time sensor-informed context-sensitive ecological momentary assessment to capture physical activity," *Journal of Medical Internet Research*, vol. 18, no. 6, p. e106, 2016.

[30] R. Ilies, N. Dimotakis, and D. Watson, "Mood, blood pressure, and heart rate at work: an experience-sampling study," *Journal of Occupational Health Psychology*, vol. 15, no. 2, pp. 120–130, 2010.

[31] V. Pejovic, N. Lathia, C. Mascolo, and M. Musolesi, “Mobile-based experience sampling for behaviour research,” in *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, and A. Košir, Eds. Cham: Springer International Publishing, 2016, pp. 141–161.

[32] M. Wichers, C. J. P. Simons, I. M. A. Kramer, J. A. Hartmann, C. Lothmann, I. Myin-Germeys, A. L. van Bemmel, F. Peeters, P. Delespaul, and J. van Os, “Momentary assessment technology as a tool to help patients with depression help themselves,” *Acta Psychiatrica Scandinavica*, vol. 124, no. 4, pp. 262–272, Oct 2011.

[33] J. van Os, S. Verhagen, A. Marsman, F. Peeters, M. Bak, M. Marcelis, M. Drukker, U. Reininghaus, N. Jacobs, T. Lataster, C. Simons, R. Lousberg, S. Gülöksüz, C. Leue, P. C. Groot, W. Viechtbauer, and P. Delespaul, “The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice,” *Depression and Anxiety*, vol. 34, no. 6, pp. 481–493, Jun 2017.

[34] P. Klasnja and W. Pratt, “Healthcare in the pocket: mapping the space of mobile-phone health interventions,” *Journal of Biomedical Informatics*, vol. 45, no. 1, pp. 184–198, 2012.

[35] M. Fuller-Tyszkiewicz, H. Skouteris, B. Richardson, J. Blore, M. Holmes, and J. Mills, “Does the burden of the experience sampling method undermine data quality in state body image research?” *Body Image*, vol. 10, no. 4, pp. 607–613, Sep 2013.

[36] V. Harrison, J. Proudfoot, P. P. Wee, G. Parker, D. H. Pavlovic, and V. Manicavasagar, “Mobile mental health: review of the emerging field and proof of concept study,” *Journal of Mental Health*, vol. 20, no. 6, pp. 509–524, 2011.

[37] A. W. Sokolovsky, R. J. Mermelstein, and D. Hedeker, “Factors predicting compliance to ecological momentary assessment among adolescent smokers,” *Nicotine and Tobacco Research*, vol. 16, no. 3, pp. 351–358, Mar 2014.

[38] F. M. Bos, F. J. Blaauw, E. Snippe, L. van der Krieke, P. de Jonge, and M. Wichers, “Exploring the emotional dynamics of subclinically depressed individuals with and without anhedonia: an experience sampling study,” *Journal of Affective Disorders*, vol. 228, pp. 186–193, Mar 2018.

[39] A. A. Stone, S. Shiffman, J. E. Schwartz, J. E. Broderick, and M. R. Hufford, “Patient compliance with paper and electronic diaries,” *Controlled Clinical Trials*, vol. 24, no. 2, pp. 182–199, 2003.

[40] R. McCarney, J. Warner, S. Iliffe, R. Van Haselen, M. Griffin, and P. Fisher, "The Hawthorne Effect: a randomised, controlled trial," *BMC Medical Research Methodology*, vol. 7, no. 1, p. 30, Dec 2007.

[41] L. A. Aaron, J. A. Turner, L. Mancl, H. Brister, and C. N. Sawchuk, "Electronic diary assessment of pain-related variables: is reactivity a problem?" *The Journal of Pain*, vol. 6, no. 2, pp. 107–115, 2005.

[42] P. R. Clifford and C. M. Davis, "Alcohol treatment research assessment exposure: a critical review of the literature," *Psychology of Addictive Behaviors*, vol. 26, no. 4, pp. 773–781, 2012.

[43] R. A. Carels, H. M. Cacciapaglia, S. Rydin, O. M. Douglass, and J. Harper, "Can social desirability interfere with success in a behavioral weight loss program?" *Psychology & Health*, vol. 21, no. 1, pp. 65–78, Feb 2006.

[44] S. A. Adams, C. E. Matthews, C. B. Ebbeling, C. G. Moore, J. E. Cunningham, J. Fulton, and J. R. Hebert, "The effect of social desirability and social approval on self-reports of physical activity," *American Journal of Epidemiology*, vol. 161, no. 4, pp. 389–398, Feb 2005.

[45] J. R. Hebert, L. Clemow, L. Pbert, I. S. Ockene, and J. K. Ockene, "Social desirability bias in dietary self-report may compromise the validity of dietary intake measures," *International Journal of Epidemiology*, vol. 24, no. 2, pp. 389–398, Apr 1995.

[46] L. F. Barrett and D. J. Barrett, "An introduction to computerized experience sampling in psychology," *Social Science Computer Review*, vol. 19, no. 2, pp. 175–185, 2001.

[47] L. Wheeler and H. T. Reis, "Self-recording of everyday life events: Origins, types, and uses," *Journal of Personality*, vol. 59, no. 3, pp. 339–354, 1991.

[48] E. E. Brannon, C. C. Cushing, C. J. Crick, and T. B. Mitchell, "The promise of wearable sensors and ecological momentary assessment measures for dynamical systems modeling in adolescents: a feasibility and acceptability study," *Translational Behavioral Medicine*, vol. 6, no. 4, pp. 558–565, Dec 2016.

[49] K. F. Burnett, C. B. Taylor, and W. S. Agras, "Ambulatory computer-assisted therapy for obesity: a new frontier for behavior therapy." *Journal of Consulting and Clinical Psychology*, vol. 53, no. 5, pp. 698–703, 1985.

[50] K. M. Eddington, C. J. Burgin, P. J. Silvia, N. Fallah, C. Majestic, and T. R. Kwapil, “The effects of psychotherapy for major depressive disorder on daily mood and functioning: a longitudinal experience sampling study,” *Cognitive Therapy and Research*, vol. 41, no. 2, pp. 266–277, Apr 2017.

[51] G. P. Galloway, R. Didier, K. Garrison, and J. Mendelson, “Feasibility of ecological momentary assessment using cellular telephones in methamphetamine dependent subjects,” *Substance Abuse*, vol. 1, pp. 9–14, Jan 2008.

[52] J. L. Pearson, H. Elmasry, B. Das, S. L. Smiley, L. F. Rubin, T. DeAtley, E. Harvey, Y. Zhou, R. Niaura, and D. B. Abrams, “Comparison of ecological momentary assessment versus direct measurement of e-cigarette use with a Bluetooth-enabled e-cigarette: a pilot study,” *JMIR Research Protocols*, vol. 6, no. 5, p. e84, May 2017.

[53] S. L. Smiley, H. Elmasry, M. Webb Hooper, R. S. Niaura, A. B. Hamilton, and N. G. Milburn, “Feasibility of ecological momentary assessment of daily sexting and substance use among young adult African American gay and bisexual men: a pilot study,” *JMIR Research Protocols*, vol. 6, no. 2, p. e9, Feb 2017.

[54] S. N. Zenk, I. Horoi, K. K. Jones, L. Finnegan, C. Corte, B. Riley, and J. Wilbur, “Environmental and personal correlates of physical activity and sedentary behavior in African American women: an ecological momentary assessment study,” *Women & Health*, vol. 57, no. 4, pp. 446–462, Apr 2017.

[55] A. Shoham, P. Goldstein, R. Oren, D. Spivak, and A. Bernstein, “Decentering in the process of cultivating mindfulness: an experience-sampling study in time and context,” *Journal of Consulting and Clinical Psychology*, vol. 85, no. 2, pp. 123–134, 2017.

[56] N. V. Berkel, D. Ferreira, and V. Kostakos, “The experience sampling method on mobile devices,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–40, 2017.

[57] S. S. Intille, “Technological innovations enabling automatic, context-sensitive ecological momentary assessment,” in *The science of real-time data capture: Self-reports in health research.* Oxford, UK: Oxford University Press, 2007, pp. 308–337.

[58] G. Hsieh, I. Li, A. Dey, J. Forlizzi, and S. E. Hudson, “Using visualizations to increase compliance in experience sampling,” in *Proceedings of the 10th international conference on Ubiquitous computing*, Seoul, Korea, 2008, pp. 164–167.

[59] N. Van Berkel, J. Goncalves, S. Hosio, and V. Kostakos, “Gamification of mobile experience sampling improves data quality and quantity,” *Proceedings of the ACM on*

Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1, no. 3, pp. 1–21, Sep 2017.

[60] A. Ponnada, C. Haynes, D. Maniar, J. Manjourides, and S. Intille, "Microinteraction ecological momentary assessment response rates: Effect of microinteractions or the smartwatch?" *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 92, 2017.

[61] R. Band, C. Barrowclough, K. Caldwell, R. Emsley, and A. Wearden, "Activity patterns in response to symptoms in patients being treated for chronic fatigue syndrome: an experience sampling methodology study." *Health Psychology*, vol. 36, no. 3, pp. 264–269, 2017.

[62] D. Kimhy, M. M. Wall, M. C. Hansen, J. Vakhrusheva, C. J. Choi, P. Delespaul, N. Tarrier, R. P. Sloan, and D. Malaspina, "Autonomic regulation and auditory hallucinations in individuals with schizophrenia: an experience sampling study," *Schizophrenia Bulletin*, vol. 43, no. 4, pp. 754–763, Jul 2017.

[63] E. Moitra, B. A. Gaudiano, C. H. Davis, and D. Ben-Zeev, "Feasibility and acceptability of post-hospitalization ecological momentary assessment in patients with psychotic-spectrum disorders," *Comprehensive Psychiatry*, vol. 74, pp. 204–213, Apr 2017.

[64] E. M. W. Tong, G. D. Bishop, H. C. Enkelmann, Y. P. Why, S. M. Diong, M. Khader, and J. Ang, "Emotion and appraisal: a study using ecological momentary assessment," *Cognition and Emotion*, vol. 21, no. 7, pp. 1361–1381, Nov 2007.

[65] P. S. Santangelo, J. Koenig, V. Funke, P. Parzer, F. Resch, U. W. Ebner-Priemer, and M. Kaess, "Ecological momentary assessment of affective and interpersonal instability in adolescent non-suicidal self-injury," *Journal of Abnormal Child Psychology*, vol. 45, no. 7, pp. 1429–1438, Oct 2017.

[66] M. aan het Rot, K. Hogenelst, and R. A. Schoevers, "Mood disorders in everyday life: a systematic review of experience sampling and ecological momentary assessment studies," *Clinical Psychology Review*, vol. 32, no. 6, pp. 510–523, 2012.

[67] M. Morren, S. Dulmen, J. Ouwerkerk, and J. Bensing, "Compliance with momentary pain measurement using electronic diaries: a systematic review," *European Journal of Pain*, vol. 13, no. 4, pp. 354–365, 2009.

[68] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *World Sensor Web Workshop, ACM Sensys 2006*, Boulder, Colorado, 2006.

[69] P. Zhou, Y. Zheng, and M. Li, "How long to wait? Predicting bus arrival time with mobile phone based participatory sensing," *IEEE Transactions on Mobile Computing*, vol. 13, no. 6, pp. 1228–1241, 2014.

[70] S. Anawar, W. A. W. Adnan, and R. Ahmad, "A design guideline for non-monetary incentive mechanics in mobile health participatory sensing system," *International Journal of Applied Engineering Research*, vol. 12, no. 21, pp. 11 039–11 049, 2017.

[71] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *The American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000.

[72] D. H. Pink, *Drive: The surprising truth about what motivates us*. New York, NY, USA: Riverhead Books, 2009.

[73] D. M. Aanensen, D. M. Huntley, E. J. Feil, F. Al-Own, and B. G. Spratt, "EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection," *PLOS One*, vol. 4, no. 9, p. e6968, Sep 2009.

[74] C. Hartung, A. Lerer, Y. Anokwa, C. Tseng, W. Brunette, and G. Borriello, "Open data kit: tools to build information services for developing regions," in *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, London, UK, 2010, p. 18.

[75] K. E. Heron and J. M. Smyth, "Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments," *British Journal of Health Psychology*, vol. 15, no. 1, pp. 1–39, 2010.

[76] D. Spruijt-Metz, C. K. Wen, G. O'Reilly, M. Li, S. Lee, B. Emken, U. Mitra, M. Annavaram, G. Ragusa, and S. Narayanan, "Innovations in the use of interactive technology to support weight management," *Current Obesity Reports*, vol. 4, no. 4, pp. 510–519, 2015.

[77] J. Bort-Roig, N. D. Gilson, A. Puig-Ribera, R. S. Contreras, and S. G. Trost, "Measuring and influencing physical activity with smartphone technology: a systematic review," *Sports Medicine*, vol. 44, no. 5, pp. 671–686, 2014.

[78] F. Naughton, S. Hopewell, N. Lathia, R. Schalbroeck, C. Brown, C. Mascolo, A. McEwen, and S. Sutton, "A context-sensing mobile phone app (Q sense) for smoking cessation: a mixed-methods study," *JMIR mHealth and uHealth*, vol. 4, no. 3, p. e106, 2016.

[79] I. Kramer, C. J. Simons, J. A. Hartmann, C. Menne-Lothmann, W. Viechtbauer, F. Peeters, K. Schruers, A. L. Bemmel, I. Myin-Germeys, P. Delespaul *et al.*, "A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial," *World Psychiatry*, vol. 13, no. 1, pp. 68–77, 2014.

[80] B. Freisthler, S. Lipperman-Kreda, M. Bersamin, and P. J. Gruenewald, "Tracking the when, where, and with whom of alcohol use: integrating ecological momentary assessment and geospatial data to examine risk for alcohol-related problems," *Alcohol Research: Current Reviews*, vol. 36, no. 1, pp. 29–38, 2014.

[81] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "EmotionSense: a mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Copenhagen, Denmark, 2010, pp. 281–290.

[82] N. Lathia, K. Rachuri, C. Mascolo, and G. Roussos, "Open source smartphone libraries for computational social science," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing adjunct publication*, Zurich, Switzerland, 2013, pp. 911–920.

[83] T. R. Kirchner and S. Shiffman, "Spatio-temporal determinants of mental health and well-being: advances in geographically-explicit ecological momentary assessment (GEMA)," *Social Psychiatry and Psychiatric Epidemiology*, vol. 51, no. 9, pp. 1211–1223, Sep 2016.

[84] J. Ho and S. S. Intille, "Using context-aware computing to reduce the perceived burden of interruptions from mobile devices," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, Portland, OR, USA, 2005, pp. 909–918.

[85] B. Poppinga, W. Heuten, and S. Boll, "Sensor-based identification of opportune moments for triggering notifications," *IEEE Pervasive Computing*, vol. 13, no. 1, pp. 22–29, 2014.

[86] N. Lathia, V. Pejovic, K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "Smartphones for large-scale behavior change interventions," *IEEE Pervasive Computing*, vol. 12, no. 3, pp. 66–73, 2013.

[87] D. Ben-Zeev, C. J. Brenner, M. Begale, J. Duffecy, D. C. Mohr, and K. T. Mueser, "Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia," *Schizophrenia Bulletin*, vol. 40, no. 6, pp. 1244–1253, Nov 2014.

[88] L. A. Jibb, B. J. Stevens, P. C. Nathan, E. Seto, J. A. Cafazzo, and J. N. Stinson, "A smartphone-based pain management app for adolescents with cancer: establishing system requirements and a pain care algorithm based on literature review, interviews, and consensus," *Journal of Medical Internet Research*, vol. 16, no. 3, p. e15, Mar 2014.

[89] E. Seto, K. J. Leonard, J. A. Cafazzo, J. Barnsley, C. Masino, and H. J. Ross, "Developing healthcare rule-based expert systems: case study of a heart failure telemonitoring system," *International Journal of Medical Informatics*, vol. 81, no. 8, pp. 556–565, Aug 2012.

[90] A. S. Juarascio, S. P. Goldstein, S. M. Manasse, E. M. Forman, and M. L. Butryn, "Perceptions of the feasibility and acceptability of a smartphone application for the treatment of binge eating disorders: qualitative feedback from a user population and clinicians," *International Journal of Medical Informatics*, vol. 84, no. 10, pp. 808–816, Oct 2015.

[91] H. Sarker, M. Sharmin, A. A. Ali, M. M. Rahman, R. Bari, S. M. Hossain, and S. Kumar, "Assessing the availability of users to engage in just-in-time intervention in the natural environment," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, WA, USA, 2014, pp. 909–920.

[92] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr, "Harnessing context sensing to develop a mobile intervention for depression," *Journal of Medical Internet Research*, vol. 13, no. 3, p. e55, Aug 2011.

[93] M. E. Hilliard, A. Hahn, A. K. Ridge, M. N. Eakin, and K. A. Riekert, "User preferences and design recommendations for an mHealth app to promote cystic fibrosis self-management," *Journal of Medical Internet Research*, vol. 16, no. 10, p. e44, Oct 2014.

[94] L. Simons, A. Z. Valentine, C. J. Falconer, M. Groom, D. Daley, M. P. Craven, Z. Young, C. Hall, and C. Hollis, "Developing mHealth remote monitoring technology for attention deficit hyperactivity disorder: a qualitative study eliciting user priorities and needs," *JMIR mHealth and uHealth*, vol. 4, no. 1, p. e31, Mar 2016.

[95] N. Ramanathan, D. Swendeman, W. S. Comulada, D. Estrin, and M. J. Rotheram-Borus, "Identifying preferences for mobile health applications for self-monitoring and self-management: Focus group findings from HIV-positive persons and young mothers," *International Journal of Medical Informatics*, vol. 82, no. 4, pp. e38–46, Apr 2013.

[96] G. F. Dunton, E. Dzubur, K. Kawabata, B. Yanez, B. Bo, and S. Intille, "Development of a smartphone application to measure physical activity using sensor-assisted self-report," *Frontiers in Public Health*, vol. 2, p. 12, 2014.

[97] J. E. Bardram, M. Frost, K. Szántó, M. Faurholt-Jepsen, M. Vinberg, and L. V. Kessing, "Designing mobile health technology for bipolar disorder: a field trial of the monarca system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, 2013, pp. 2627–2636.

[98] S. W. Miyamoto, S. Henderson, H. M. Young, A. Pande, and J. J. Han, "Tracking health data is not enough: A qualitative exploration of the role of healthcare partnerships and mHealth technology to promote physical activity and to sustain behavior change," *JMIR mHealth and uHealth*, vol. 4, no. 1, p. e5, Jan 2016.

[99] D. K. King, D. J. Toobert, J. D. Portz, L. A. Strycker, A. Doty, C. Martin, J. M. Boggs, A. J. Faber, C. R. Geno, and R. E. Glasgow, "What patients want: relevant health information technology for diabetes self-management," *Health and Technology*, vol. 2, no. 3, pp. 147–157, Sep 2012.

[100] J. A. Naslund, K. A. Aschbrenner, and S. J. Bartels, "Wearable devices and smartphones for activity tracking among people with serious mental illness," *Mental Health and Physical Activity*, vol. 10, pp. 10–17, Mar 2016.

[101] T. Ludwig, J. Dax, V. Pipek, and D. Randall, "Work or leisure? Designing a user-centered approach for researching activity "in the wild"," *Personal and Ubiquitous Computing*, vol. 20, no. 4, pp. 487–515, 2016.

[102] L. Dennison, L. Morrison, G. Conway, and L. Yardley, "Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study." *Journal of Medical Internet Research*, vol. 15, no. 4, p. e86, Apr 2013.

[103] P. Markopoulos, N. Batalas, and A. Timmermans, "On the use of personalization to enhance compliance in experience sampling," in *Proceedings of the European Conference on Cognitive Ergonomics 2015*, Warsaw, Poland, 2015, pp. 1–4.

[104] S. Vhaduri and C. Poellabauer, "Human factors in the design of longitudinal smartphone-based wellness surveys," in *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, Chicago, IL, USA, 2016, pp. 156–167.

[105] K. Shilton, N. Ramanathan, S. Reddy, V. Samanta, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Participatory design of sensing networks: strengths and challenges,"

in *Proceedings of the Tenth Anniversary Conference on Participatory Design 2008*, Bloomington, IN, USA, 2008, pp. 282–285.

[106] S. Carter, J. Mankoff, and J. Heer, "Momento: support for situated ubicomp experimentation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, CA, USA, 2007, pp. 125–134.

[107] S. C. Reid, S. D. Kauer, A. S. Khor, S. J. C. Hearps, L. A. Sanci, A. D. Kennedy, and G. C. Patton, "Using a mobile phone application in youth mental health: an evaluation study," *Australian Family Physician*, vol. 41, no. 9, pp. 711–714, Dec 2012.

[108] W. T. Riley, D. E. Rivera, A. A. Atienza, W. Nilsen, S. M. Allison, and R. Mermelstein, "Health behavior models in the age of mobile interventions: are our theories up to the task?" *Translational behavioral medicine*, vol. 1, no. 1, pp. 53–71, 2011.

[109] T. McCurdie, S. Taneva, M. Casselman, M. Yeung, C. McDaniel, W. Ho, and J. Cafazzo, "mHealth consumer apps: the case for user-centered design," *Biomedical Instrumentation & Technology*, vol. 46, no. s2, pp. 49–56, 2012.

[110] D. Swendeman, S. Farmer, D. Mindry, S.-J. Lee, and M. Medich, "HIV care providers' attitudes regarding mobile phone applications and web-based dashboards to support patient self-management and care coordination: results from a qualitative feasibility study," *Journal of HIV and AIDS*, vol. 2, no. 4, pp. 1310–1318, Oct 2016.

[111] C. Scaffidi, M. Shaw, and B. Myers, "Estimating the numbers of end users and end user programmers," in *Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on*, Dallas, TX, USA, Sep 2005, pp. 207–214.

[112] C. Kelleher and R. Pausch, "Lowering the barriers to programming: a taxonomy of programming environments and languages for novice programmers," *ACM Computing Surveys (CSUR)*, vol. 37, no. 2, pp. 83–137, 2005.

[113] J. Segal, "Some problems of professional end user developers," in *Visual Languages and Human-Centric Computing, 2007. VL/HCC 2007. IEEE Symposium on*, Coeur d'Alene, ID, USA, 2007, pp. 111–118.

[114] A. Henderson and M. Kyng, "There's no place like home: Continuing design in use," in *Human-computer interaction*, R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 793–803.

[115] A. Mørch, “Three levels of end-user tailoring: customization, integration, and extension,” in *Computers and Design in Context*, M. Kyng and L. Mathiassen, Eds. MIT Press, Cambridge, MA, USA, 1997, pp. 51–76.

[116] M. Spahn, C. Dörner, and V. Wulf, “End user development: approaches towards a flexible software design,” in *European Conference on Information Systems (ECIS)*, 2008, pp. 303–314.

[117] B. A. Nardi, *A Small Matter of Programming: Perspectives on End User Computing*, 1st ed. Cambridge, MA, USA: MIT Press, 1993.

[118] G. Fischer, D. Fogli, and A. Piccinno, “Revisiting and broadening the meta-design framework for end-user development,” in *New Perspectives in End-User Development*, F. Paternò and V. Wulf, Eds. Cham: Springer International Publishing, 2017, pp. 61–97.

[119] A. Cypher and D. C. Halbert, *Watch what I do: programming by demonstration*. Cambridge, MA, USA: MIT press, 1993.

[120] A. K. Dey, R. Hamid, C. Beckmann, I. Li, and D. Hsu, “a CAPpella: programming by demonstration of context-aware applications,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria, 2004, pp. 33–40.

[121] J. F. Pane, B. A. Myers, and L. B. Miller, “Using HCI techniques to design a more usable programming system,” in *Human Centric Computing Languages and Environments, 2002. Proceedings. IEEE 2002 Symposia on*, Arlington, VA, USA, 2002, pp. 198–206.

[122] H. Liu and H. Lieberman, “Programmatic semantics for natural language interfaces,” in *CHI’05 Extended Abstracts on Human Factors in Computing Systems*, Portland, OR, USA, 2005, pp. 1597–1600.

[123] M. M. Burnett and C. Scaffidi, “End-user development,” in *Encyclopedia of Human-Computer Interaction*. The Interaction Design Foundation, 2011, accessed 17th March 2018. [Online]. Available: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/end-user-development

[124] D. Fogli and L. P. Provenza, “A meta-design approach to the development of e-government services,” *Journal of Visual Languages & Computing*, vol. 23, no. 2, pp. 47–62, 2012.

[125] G. Desolda, C. Ardito, and M. Matera, “Empowering end users to customize their smart environments: Model, composition paradigms, and domain-specific tools,” *ACM Transactions on Computer-Human Interaction*, vol. 24, no. 2, pp. 12:1–12:52, Apr. 2017.

[126] C. Elliott, V. Vijayakumar, W. Zink, and R. Hansen, "National instruments LabVIEW: a programming environment for laboratory automation and measurement," *JALA: Journal of the Association for Laboratory Automation*, vol. 12, no. 1, pp. 17–24, 2007.

[127] M. Burnett, "Visual programming," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, J. G. Webster, Ed. New York, NY, USA: John Wiley & Sons, Inc., Dec 1999, vol. 32, no. 1-3, pp. 275–283.

[128] K. N. Whitley, "Visual programming languages and the empirical evidence for and against," *Journal of Visual Languages & Computing*, vol. 8, no. 1, pp. 109–142, 1997.

[129] J. Maloney, M. Resnick, N. Rusk, B. Silverman, and E. Eastmond, "The Scratch programming language and environment," *ACM Transactions on Computing Education (TOCE)*, vol. 10, no. 4, p. 16, 2010.

[130] S. Cooper, W. Dann, and R. Pausch, "Alice: a 3-D tool for introductory programming concepts," *Journal of Computing Sciences in Colleges*, vol. 15, pp. 107–116, 2000.

[131] D. Wolber, H. Abelson, E. Spertus, and L. Looney, *App Inventor: Create Your Own Android Apps*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2011.

[132] B. A. Myers, "Taxonomies of visual programming and program visualization," *Journal of Visual Languages and Computing*, vol. 1, no. 1, pp. 97–123, Mar 1990.

[133] M. M. Burnett and M. J. Baker, "A classification system for visual programming languages," *Journal of Visual Languages and Computing*, vol. 5, no. 3, pp. 287–300, 1994.

[134] G. Costagliola, A. Delucia, S. Orefice, and G. Polese, "A classification framework to support the design of visual languages," *Journal of Visual Languages & Computing*, vol. 13, no. 6, pp. 573–600, 2002.

[135] M. C. Carlisle, "Raptor: a visual programming environment for teaching object-oriented programming," *Journal of Computing Sciences in Colleges*, vol. 24, no. 4, pp. 275–281, Apr 2009.

[136] D. S. Touretzky, "Teaching Kodu with physical manipulatives," *ACM Inroads*, vol. 5, no. 4, pp. 44–51, 2014.

[137] J. Lee, L. Garduño, E. Walker, and W. Burleson, "A tangible programming tool for creation of context-aware applications," in *Proceedings of the 2013 ACM International*

Joint conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 2013, pp. 391–400.

[138] D. Bau, J. Gray, C. Kelleher, J. Sheldon, and F. Turbak, "Learnable programming: blocks and beyond," *Communications of the ACM*, vol. 60, no. 6, pp. 72–80, May 2017.

[139] G. Fischer, E. Giaccardi, Y. Ye, A. G. Sutcliffe, and N. Mehandjiev, "Meta-design: A manifesto for end-user development," *Communications of the ACM*, vol. 47, no. 9, pp. 33–37, 2004.

[140] M. F. Costabile, D. Fogli, P. Mussio, and A. Piccinno, "Visual interactive systems for end-user development: a model-based design methodology," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, vol. 37, no. 6, pp. 1029–1046, Nov 2007.

[141] A. J. Ko, B. Myers, M. B. Rosson, G. Rothermel, M. Shaw, S. Wiedenbeck, R. Abraham, L. Beckwith, A. Blackwell, M. Burnett, M. Erwig, C. Scaffidi, J. Lawrance, and H. Lieberman, "The state of the art in end-user software engineering," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–44, 2011.

[142] F. Cabitza, D. Fogli, and A. Piccinno, ""Each to his own": distinguishing activities, roles and artifacts in EUD practices," in *Smart Organizations and Smart Artifacts*, L. Caporarello, B. Di Martino, and M. Martinez, Eds. Cham: Springer International Publishing, 2014, pp. 193–205.

[143] S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, Florence, Italy, 2008, pp. 111–120.

[144] A. Namoun, A. Daskalopoulou, N. Mehandjiev, and Z. Xun, "Exploring mobile end user development: existing use and design factors," *IEEE Transactions on Software Engineering*, vol. 42, no. 10, pp. 960–976, 2016.

[145] V. Venkatesh and F. D. Davis, "A theoretical extension of the Technology Acceptance Model: four longitudinal field studies," *Management Science*, vol. 46, no. 2, pp. 186–204, Feb 2000.

[146] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, Sep. 2003.

[147] T. Conner, "Experience sampling and ecological momentary assessment with mobile phones," 2015, accessed 11th March 2018. [Online]. Available: http://www.otago.ac.nz/psychology/otago047475.pdf

[148] S. S. Intille, E. M. Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, L. Bao, and K. Larson, "Tools for studying behavior and technology in natural settings," in *UbiComp 2003: Ubiquitous Computing*, A. K. Dey, A. Schmidt, and J. F. McCarthy, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 157–174.

[149] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, "MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones," in *Proceedings of the 5th international conference on Mobile systems, applications and services*, San Juan, Puerto Rico, 2007, pp. 57–70.

[150] A. Gaggioli, G. Pioggia, G. Tartarisco, G. Baldus, D. Corda, P. Cipresso, and G. Riva, "A mobile data collection platform for mental health research," *Personal and Ubiquitous Computing*, vol. 17, no. 2, pp. 241–251, 2013.

[151] D. Ferreira, V. Kostakos, and A. K. Dey, "AWARE: mobile context instrumentation framework," *Frontiers in ICT*, vol. 2, p. 6, 2015.

[152] H. Tangmunarunkit, C.-K. Hsieh, B. Longstaff, S. Nolen, J. Jenkins, C. Ketcham, J. Selsky, F. Alquaddoomi, D. George, J. Kang *et al.*, "Ohmage: a general and extensible end-to-end participatory sensing platform," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 1–21, 2015.

[153] H. Xiong, Y. Huang, L. E. Barnes, and M. S. Gerber, "Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Heidelberg, Germany, 2016, pp. 415–426.

[154] A. K. Dey, T. Sohn, S. Streng, and J. Kodama, "iCAP: interactive prototyping of context-aware applications," in *Pervasive Computing*, K. P. Fishkin, B. Schiele, P. Nixon, and A. Quigley, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 254–271.

[155] P. Korpipaa, E.-J. Malm, T. Rantakokko, V. Kyllonen, J. Kela, J. Mantyjarvi, J. Hakkila, and I. Kansala, "Customizing user interaction in smart phones," *IEEE Pervasive Computing*, vol. 5, no. 3, pp. 82–90, Jul. 2006.

[156] J. Fischer, "Experience-sampling tools: a critical review," in *Mobile Living Labs 09: Methods and Tools for Evaluation in the Wild, part of MobileHCI 2009*, Bonn, Germany, 2009, pp. 1–3.

[157] V. Pejovic and M. Musolesi, "InterruptMe: designing intelligent prompting mechanisms for pervasive applications," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Seattle, WA, USA, 2014, pp. 897–908.

[158] D. Fogli and A. Piccinno, "Co-evolution of end-user developers and systems in multi-tiered proxy design problems," in *End-User Development*, Y. Dittrich, M. Burnett, A. Mørch, and D. Redmiles, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 153–168.

[159] B. Ur, E. McManus, M. Pak Yong Ho, and M. L. Littman, "Practical trigger-action programming in the smart home," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Toronto, Ontario, Canada, 2014, pp. 803–812.

[160] J. E. Palmier-Claus, I. Myin-Germeys, E. Barkus, L. Bentley, A. Udachina, P. A. Delespaul, S. W. Lewis, and G. Dunn, "Experience sampling research in individuals with mental illness: reflections and guidance," *Acta Psychiatrica Scandinavica*, vol. 123, no. 1, pp. 12–20, Jan 2011.

[161] T. R. G. Green and M. Petre, "Usability analysis of visual programming environments: a 'cognitive dimensions' framework," *Journal of Visual Languages & Computing*, vol. 7, no. 2, pp. 131–174, 1996.

[162] D. Moody, "The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering," *IEEE Transactions on Software Engineering*, vol. 35, no. 6, pp. 756–779, Nov 2009.

[163] T. R. G. Green, "Cognitive dimensions of notations," in *People and Computers V*, A. Sutcliffe and L. Macaulay, Eds. Cambridge, UK: Cambridge University Press, 1989, pp. 443–460.

[164] S. Turkle and S. Papert, "Epistemological pluralism: styles and voices within the computer culture," *Signs: Journal of Women in Culture and Society*, vol. 16, no. 1, pp. 128–157, 1990.

[165] J. Bertin, *Semiology of graphics: diagrams, networks, maps*. Madison, WI, USA: University of Wisconsin press, 1983.

[166] T. Green and A. Blackwell, "Cognitive dimensions of information artefacts: a tutorial," 1998, accessed 17th March 2018. [Online]. Available: https://www.cl.cam.ac.uk/~afb21/CognitiveDimensions/CDtutorial.pdf

[167] M. García-Herranz, P. Haya, and X. Alamán, "Towards a ubiquitous end–user programming system for smart spaces," *Journal of Universal Computer Science*, vol. 16, no. 12, pp. 1633–1649, 2010.

[168] Y. Dahl and R.-M. Svendsen, "End-user composition interfaces for smart environments: A preliminary study of usability factors," in *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, A. Marcus, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 118–127.

[169] J. H. Larkin and H. A. Simon, "Why a diagram is (sometimes) worth ten thousand words," *Cognitive Science*, vol. 11, no. 1, pp. 65–100, 1987.

[170] A. Paivio, "Dual coding theory: retrospect and current status," *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 45, no. 3, pp. 255–287, 1991.

[171] G. A. Miller, "The magic number seven, plus or minus two: some limits on our capacity for processing information." *Psychological Review*, vol. 63, pp. 81–97, 1956.

[172] N. Lathia, K. K. Rachuri, C. Mascolo, and P. J. Rentfrow, "Contextual dissonance: design bias in sensor-based experience sampling methods," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Zurich, Switzerland, 2013, pp. 183–192.

[173] R. Molich and J. Nielsen, "Improving a human-computer dialogue," *Communications of the ACM*, vol. 33, no. 3, pp. 338–348, Mar. 1990.

[174] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, Seattle, WA, USA, 1990, pp. 249–256.

[175] N. Bevan, "International standards for HCI and usability," *International Journal of Human-Computer Studies*, vol. 55, no. 4, pp. 533 – 552, 2001.

[176] K. Hornbæk, "Current practice in measuring usability: challenges to usability studies and research," *International Journal of Human-Computer Studies*, vol. 64, no. 2, pp. 79–102, 2006.

[177] J. Brooke, "SUS-a quick and dirty usability scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.

[178] R. A. Virzi, "Refining the test phase of usability evaluation: How many subjects is enough?" *Human Factors*, vol. 34, no. 4, pp. 457–468, 1992.

[179] J. R. Lewis, "Sample sizes for usability studies: Additional considerations," *Human Factors*, vol. 36, no. 2, pp. 368–378, 1994.

[180] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, Jul 2008.

[181] J. R. Lewis and J. Sauro, "The factor structure of the system usability scale," in *Human Centered Design*, M. Kurosu, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 94–103.

[182] O. Meerbaum-Salant, M. Armoni, and M. Ben-Ari, "Habits of programming in Scratch," in *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, Darmstadt, Germany, 2011, pp. 168–172.

[183] G. Lucci and F. Paternò, "Understanding end-user development of context-dependent applications in smartphones," in *Human-Centered Software Engineering*, S. Sauer, C. Bogdan, P. Forbrig, R. Bernhaupt, and M. Winckler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 182–198.

[184] M. P. G. Broen, V. A. M. Marsman, M. L. Kuijf, R. J. Van Oostenbrugge, J. van Os, and A. F. G. Leentjens, "Unraveling the relationship between motor symptoms, affective states and contextual factors in Parkinson's disease: a feasibility study of the experience sampling method," *PLOS One*, vol. 11, no. 3, p. e0151195, Mar 2016.

[185] T. A. Pickering, J. Huh, S. Intille, Y. Liao, M. A. Pentz, and G. F. Dunton, "Physical activity and variation in momentary behavioral cognitions: an ecological momentary assessment study," *Journal of Physical Activity and Health*, vol. 13, no. 3, pp. 344–351, Mar 2016.

[186] M. A. Russell, L. Wang, and C. L. Odgers, "Witnessing substance use increases same-day antisocial behavior among at-risk adolescents: Gene–environment interaction in a 30-day ecological momentary assessment study," *Development and psychopathology*, vol. 28, no. 4, pp. 1441–1456, 2016.

[187] R. M. Schuster, R. J. Mermelstein, and D. Hedeker, “Ecological momentary assessment of working memory under conditions of simultaneous marijuana and tobacco use,” *Addiction*, vol. 111, no. 8, pp. 1466–1476, Aug 2016.

[188] J. A. Wonderlich, J. M. Lavender, S. A. Wonderlich, C. B. Peterson, S. J. Crow, S. G. Engel, D. Le Grange, J. E. Mitchell, and R. D. Crosby, “Examining convergence of retrospective and ecological momentary assessment measures of negative affect and eating disorder behaviors,” *International Journal of Eating Disorders*, vol. 48, no. 3, pp. 305–311, Apr 2015.

[189] B. H. W. Te Lindert, J. Itzhacki, W. P. van der Meijden, M. L. Kringelbach, J. Mendoza, and E. J. W. Van Someren, “Bright environmental light ameliorates deficient subjective liking in insomnia: an experience sampling study.” *Sleep*, vol. 41, no. 4, pp. 1–13, 2018.

[190] A. Schöndube, A. Bertrams, G. Sudeck, and R. Fuchs, “Self-control strength and physical exercise: An ecological momentary assessment study,” *Psychology of Sport and Exercise*, vol. 29, pp. 19–26, Mar 2017.

[191] C. Lewis, P. G. Polson, C. Wharton, and J. Rieman, “Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1990, pp. 235–242.

[192] J. Nielsen, *Usability Engineering*. Boston, MA, USA: AP Professional, 1993.

[193] D. Armstrong, A. Gosling, J. Weinman, and T. Marteau, “The place of inter-rater reliability in qualitative research: An empirical study,” *Sociology*, vol. 31, no. 3, pp. 597–606, 1997.

[194] A. J. Ko, B. A. Myers, and H. H. Aung, “Six learning barriers in end-user programming systems,” in *Visual Languages and Human Centric Computing, 2004 IEEE Symposium on*, Rome, Italy, 2004, pp. 199–206.

[195] D. Norman, *The design of everyday things: Revised and expanded edition*. New York, NY, USA: Basic Books, 2013.

[196] W. Haines, M. Gervasio, A. Spaulding, and B. Peintner, “Recommendations for end-user development,” in *ACM Workshop on User-Centric Evaluation of Recommender Systems and their Interfaces*, 2010.

[197] J. Huang and M. Cakmak, “Supporting mental model accuracy in trigger-action programming,” *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp ’15*, pp. 215–225, 2015.

[198] A. Cockburn and A. Firth, "Improving the acquisition of small targets," in *People and Computers XVII — Designing for Society*, E. O'Neill, P. Palanque, and P. Johnson, Eds. London, UK: Springer London, 2004, pp. 181–196.

[199] N. Tillmann, M. Moskal, J. de Halleux, and M. Fahndrich, "TouchDevelop: programming cloud-connected mobile devices via touchscreen," in *Proceedings of the 10th SIGPLAN symposium on New ideas, new paradigms, and reflections on programming and software*, Portland, OR, USA, 2011, pp. 49–60.

[200] M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman *et al.*, "Scratch: programming for all," *Communications of the ACM*, vol. 52, no. 11, pp. 60–67, 2009.

[201] R. Karkar, J. Schroeder, D. A. Epstein, L. R. Pina, J. Scofield, J. Fogarty, J. A. Kientz, S. A. Munson, R. Vilardaga, and J. Zia, "Tummytrials: a feasibility study of using self-experimentation to detect individualized food triggers," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver, CO, USA, 2017, pp. 6850–6863.

[202] A. Repenning and A. Ioannidou, "What makes end-user development tick? 13 design guidelines," in *End User Development*, H. Lieberman, F. Paternò, and V. Wulf, Eds. Dordrecht: Springer Netherlands, 2006, pp. 51–85.

[203] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen, "Getting the right design and the design right," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006, pp. 1243–1252.

[204] D. A. Schon, *Educating the Reflective Practitioner. Toward a New Design for Teaching and Learning in the Professions. The Jossey-Bass Higher Education Series.* ERIC, 1987.

[205] N. Mehandjiev, A. Sutcliffe, and D. Lee, "Organizational view of end-user development," in *End User Development*, H. Lieberman, F. Paternò, and V. Wulf, Eds. Dordrecht: Springer Netherlands, 2006, pp. 371–399.

[206] R. K. Yin, *Applications of Case Study Research*, 3rd ed. Thousand Oaks, CA, USA: Sage, 2011.

[207] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, 2009.

[208] J. Lazar, J. H. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction*, 2nd ed. Cambridge, MA, USA: Morgan Kaufmann, 2017.

[209] J. Blomberg, J. Giacomi, A. Mosher, and P. Swenton-Wall, "Ethnographic field methods and their relation to design," in *Participatory Design: Principles and Practices*, D. Schuler and A. Namioka, Eds. Hillsdale, NJ, USA: Lawrence Erlbaum, 1993, pp. 123–155.

[210] E. M. Rogers, *Diffusion of Innovations*, 4th ed. New York, NY, USA: Simon and Schuster, 1995.

[211] A. F. Blackwell, "First steps in programming: a rationale for attention investment models," in *Human Centric Computing Languages and Environments, 2002. Proceedings. IEEE 2002 Symposia on*, Arlington, VA, USA, 2002, pp. 2–10.

[212] D. Tetteroo and P. Markopoulos, "EUD survival "in the wild": Evaluation challenges for field deployments and how to address them," in *New Perspectives in End-User Development*, F. Paternò and V. Wulf, Eds. Cham: Springer International Publishing, 2017, pp. 207–229.

[213] N. Batalas, "EMA IDEs: A Challenge for End User Development," in *End-User Development*, P. Díaz, V. Pipek, C. Ardito, C. Jensen, I. Aedo, and A. Boden, Eds. Cham: Springer International Publishing, 2015, pp. 259–263.

[214] N. C. Stott and R. H. Davis, "The exceptional potential in each primary care consultation," *The Journal of the Royal College of General Practitioners*, vol. 29, no. 201, pp. 201–205, 1979.

[215] P. Dourish, "Implications for design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006, pp. 541–550.

[216] M. Beaudouin-Lafon and W. Mackay, "Prototyping tools and techniques," in *The Human-computer Interaction Handbook*, J. A. Jacko and A. Sears, Eds. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2003, pp. 1006–1031.