Making Sense of Long-Term Physical Activity Tracker Data:
The challenge of Incompleteness

A thesis submitted to fulfil requirements for the
degree of Doctor of Philosophy in the School of Information Technologies at
The University of Sydney

Lie Ming Tang
June 2019

Declaration

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I, LIE MING TANG, certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged. Co-authorship declaration statements have been included in the appendix for each co-author and corresponding publications.

Date

Signature

Abstract

Millions of people have already collected weeks, months and even years of data about their own health and physical activity levels. This is expected to grow with the rise of tracking in non-dedicated devices such as smart phones and smart watches. The potential is enormous for use in personal applications as well as for public health analysis of large populations at low cost. However, beneath the promise of big data and its assumed usefulness, the reality is many people fail to wear their tracker and record data all day every day especially over the long-term. This presents a key barrier of data incompleteness. The resulting incompleteness poses an important challenge for interpreting long-term tracker data, in terms of both making sense of it and in dealing with the uncertainty of inferences based on it. Indeed, studies have shown that very few trackers made use of their own long-term data including those who have already amassed months and even years of it. While incompleteness in physical activity tracker data may appear uncontroversial, surprisingly, there has been little work into defining the problem, its extent and how it should be measured and addressed.

This thesis tackles this key challenge that is central to harnessing long-term physical activity data. We demonstrate the need for a term to describe and quantify this challenge. We introduce the term, ***adherence***, which quantifies the completeness in such data. We also offer interface designs that accounted for adherence to support self-monitoring and reflection. Bringing these together, we offer broader definitions and guidelines for incorporating adherence when making sense of long-term physical activity tracker data, both in personal applications and in public health research results.

This thesis is based on three studies. First is a semester-long study of tracker use by 237 University students. Second is a study of 21 existing long-term physical activity trackers and provided the first richly qualitative exploration of physical activity and adherence of such users. It also evaluated the *iStuckWithIt*, a long-term physical activity data user interface, and reported on insights gained within and as aided by a tutorial

and reflection scaffolding. In the final study, we drew on 12 diverse datasets, for 753 users, with over 77,000 days with data and 73,000 days missing to explore the impact of different definitions of adherence and methods for dealing with its implications.

I Dedicate This Thesis To My Beloved Mother Qiu Li Juan And My Father Qingliang Tang For Their Endless Support.

我将这篇论文献给我亲爱的母亲邱丽娟和我的父亲唐清亮为他们的无尽支持.

Acknowledgements

I would like to express my deepest gratitude to Professor Judy Kay, my research supervisor, for her ideas, patient guidance, encouragement and insightful critiques of this research work. Without whom the accomplishments of this work are inconceivable. She helped me grow as a researcher and more importantly as a thinker. Special thanks to Professor Bob Kummerfeld for his co-supervision and Professor Kalina Yacef for her advice and generous support over the years.

I have had the privilege to work alongside many outstanding researchers and wonderful people over the course of this work. I thank my collaborators at the Charles Perkins Centre. In particular, I am grateful to Professor Adrian Bauman, Lina Engelen and Professor Margaret Allman-Farinelli for their insights and advise. Thanks to Kevin Bragg, Dr Daniel Epstein, Dr Jochen Meyer who generously shared their research data, ideas and expertise. I am grateful for the help from Professor Philip Poronnik and Margot Day who dedicated their lives to the education, health and well-being of their students. I thank my fellow students for the stimulating discussions, help and collaboration over the years.

Finally, I thank all the people who participated in our studies, who donated their time and personal data to our research.

List of Publications

Publications included in this thesis.

Tang, L. M. and Kay, J. (2016). Daily & hourly adherence : towards understanding activity tracker accuracy. *CHI '16 Extended Abstracts on Human Factors in Computing Systems*

Tang, L. M. and Kay, J. (2017). Harnessing Long Term Physical Activity Data—How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):26

Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., and Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):37

Tang, L. M. and Kay, J. (2018a). Scaffolding for an olm for long-term physical activity goals. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 147–156. ACM

Tang, L. M. and Kay, J. (2018b). Understanding physical activity tracking data: wear-time matters. *Pending Submission*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aim of this thesis was to help people make sense of long-term physical activity tracker data, at both the individual and aggregate or population levels. The key to achieving this is tackling the challenge of data incompleteness.

1.1 Motivation

If an individual can maintain healthy habits, including maintaining regular levels of physical activity, they can live longer, be healthier and retard the aging process [Bandura, 2005a]. However, this can be challenging as it requires consistent and repeated effort over long periods of time. A large body of work shows that we can help people achieve this goal by improving their ability to self-regulate, including self-monitoring, self-reflection, goal setting and planning [Shilts et al., 2004, Consolvo et al., 2009, Strecher et al., 1995]. These topics have been of intense interest in the field of personal informatics, which focuses on the study of self-tracking tools and applications that support self-monitoring and reflection towards personal goals [Li et al., 2010, Rooksby et al., 2014, Choe et al., 2014, Consolvo et al., 2014, Fritz et al., 2014].

Self-monitoring refers to when individuals tracking their own data to observe actual behaviours and patterns [Bravata et al., 2007]. Indeed, many people are already using tracker data to monitor a variety of health-related personal behaviours such as food consumption, sleep and physical activity [Li et al., 2011, Rooksby et al., 2014, Choe et al., 2014, Consolvo et al., 2014, Epstein et al., 2016a]. Moreover, interviews and surveys of early adopters of this technology show that health and well-being is a key motivation for tracking. Researchers have reported that people often use tracker data to reach a goal, such as a weight loss goal, to manage medical conditions or to make better health decisions [Choe et al., 2014, Li et al., 2011, Rooksby et al., 2014].

Self-reflection refers to when individuals consider their own data and its implications for achieving their goals. Studies have shown that personal tracker data has the potential to help individuals consider how their environmental factors, such as work, school, seasons and where they live, affected their activity levels [Consolvo et al., 2014, Choe et al., 2014, Rooksby et al., 2014, Li et al., 2012, Bentley et al., 2013].

Goal setting and planning refers to plans created and actions that are performed in response to self-monitoring and reflection. For example, a person may consider lowering their goal or target from 10,000 steps per day to a more achievable 6,000 steps. Self-reflection may also prompt an individual to make plans to alter their environments, such as choosing a home that is closer to public transport or planning to be more active on weekends. There is a large body of literature that points to the potential for using goal setting as a strategy for healthy behaviour change [Strecher et al., 1995, Shilts et al., 2004, Consolvo et al., 2009]. The ability to set the right goals and plans is critical to successfully achieving goals.

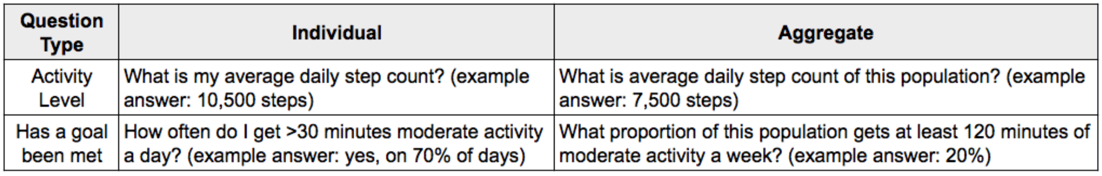
Health and well-being demand long-term goals that require consistent and repeated effort over the long-term. Long-term data has the potential to play an important role in supporting people in achieving such goals. For example, Li et al. [Li et al., 2012] found that long-term trackers wanted to be able to track data over seasons and even years. They wanted to relate their activity to contextual information in order to help them to make sense of their data. Another potential use for long-term data is to help people formulate new hypotheses and to evaluate the effectiveness of past long-term strategies. This may even help individuals to make long-term decisions such as where to live and work, e.g., *"I thought moving to the suburbs would allow me to take public transport more often, thus enabling me to be more active. Has this been true over the last two years, since I moved?"*.

Having access to fine-grained and long-term data can also allow systems and applications to draw on insights gained from a user's past in order to provide personalised and actionable recommendations. The emergence of consumer physical activity trackers such as the Fitbit has drastically improved people's ability to collect long-term physical activity data. Such trackers can provide fine-grained (e.g., per minute), objective measures (e.g., steps, active minutes) at a low cost, and these devices are designed to be worn for extended periods of time. However, there has been very little research on the use of long-term physical activity data and there has been few studies of systems and user interface designs that help people make good use of this data. Indeed, studies of existing personal trackers have reported that people make very little use of their long-term physical activity data [Bentley et al., 2013]. Interestingly, studies suggest that people have been reported to view their long-term data as valuable even though they currently have not found a use for it [Barua et al., 2013, Elsden et al., 2015].

The limited use and insights into how long-term physical activity data can be effectively used are not surprising since consumer activity trackers have only recently emerged, and at this point in time, the only very long-term users are the early adopters of this technology. As personal tracking matures and the collection of long-term data reaches the mainstream, the need for tools and user interfaces that support people in making sense of long-term data will become critical [Li et al., 2011, Consolvo et al., 2014].

When personally collected long-term physical activity data is aggregated into data collections that cover whole cohorts and populations, there is enormous potential to provide valuable insights into these cohorts or populations. Table 1.1 [1] provides examples of two types of questions that long-term physical activity data can answer with individual and aggregate level data: activity level and goal met. With their own individual long-term data, users can ask how active they have been (e.g., 10,500 steps a day over the last six months) or how often they have met their active minutes goal (e.g., at least 30 active minutes on 70% of days over the last six months). When this data is aggregated over a comparable sample, an individual can ask questions about how they

Table 1.1: Examples of important questions long-term physical activity data can answer, at the individual or aggregate level.



Question Type	Individual	Aggregate
Activity Level	What is my average daily step count? (example answer: 10,500 steps)	What is average daily step count of this population? (example answer: 7,500 steps)
Has a goal been met	How often do I get >30 minutes moderate activity a day? (example answer: yes, on 70% of days)	What proportion of this population gets at least 120 minutes of moderate activity a week? (example answer: 20%)

[1] Table 1.1 also appear as Table 5.1.

compare to similar individuals.

Aggregate or population data is also useful for public health data analysis. For example, public health studies can examine the activity level of a population or cohort of people (e.g., only 20% of people achieve at least 120 active minutes per week). This can then be combined with other health and social factors to better inform future health recommendations and public policies. For example, Althoff et al. [Althoff et al., 2017] demonstrated the public health use of large and long-term aggregate data in a study comparing activity inequality between people in 111 countries.

The Challenge of Incompleteness

A key challenge to realising the full potential of long-term physical activity data is that this data is often incomplete. Incompleteness can be due to different wearing habits or preferences (e.g., a person may only wear their tracker during part of a day, weekdays, weekends, or may have stopped wearing it when the tracker did not match the fashion) [Shih et al., 2015, Meyer et al., 2017, Harrison et al., 2015], the device or design of the device (e.g., battery life, no waterproofing) or physical constraints (e.g., workplace or sports safety requirements, clothing) [Epstein et al., 2016b, Fritz et al., 2014, Shih et al., 2015]. Inconsistent gaps in data of varying lengths can also occur due to other reasons, such as changing of devices (e.g., lost, broken, new), going on holiday, the presence of an injury or health problems or because the user may have simply lost interest in tracking for a while [Epstein et al., 2016b]. Indeed, studies of long-term physical activity trackers have found that incompleteness is far more common than completeness [Epstein et al., 2016b, Meyer et al., 2017].

Incomplete data can be a threat to reports and systems that depend on it. Fogg [Fogg, 2003] warned that systems that produce inaccurate or questionable data can result in loss of trust, thus limiting their usefulness. Studies of personal informatics and healthy behaviour change systems have produced similar observations [Consolvo et al., 2014, Bentley et al., 2013].

One approach to addressing the problem of incomplete tracker data is to find a way to encourage or help people to be more consistent in wearing their tracker [Tudor-Locke et al., 2015, Faust et al., 2017, Gouveia et al., 2015].

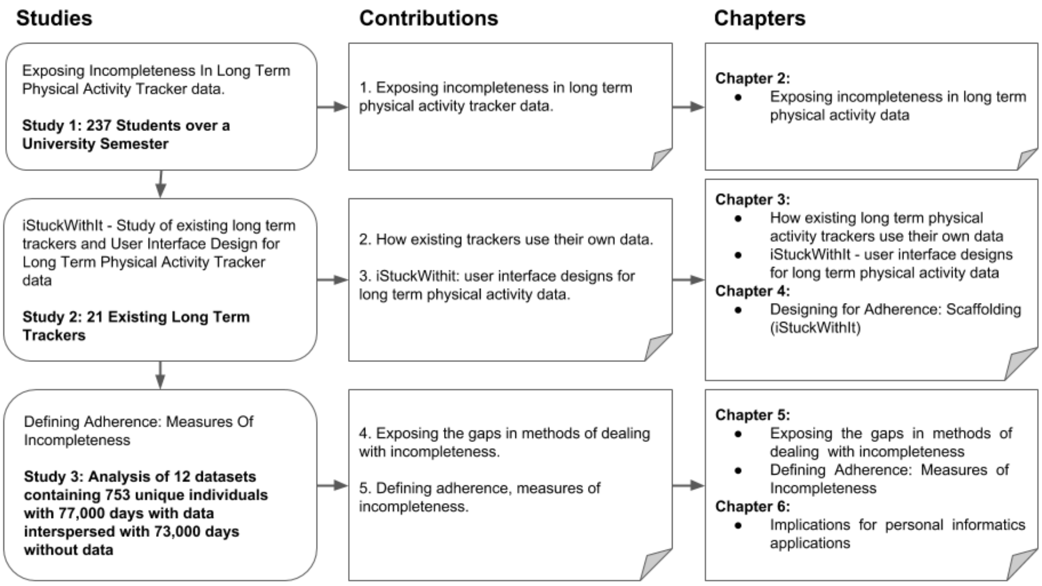

Studies
Contributions
Chapters
Exposing Incompleteness In Long Term Physical Activity Tracker data.
Study 1: 237 Students over a University Semester
1. Exposing incompleteness in long term physical activity tracker data.
Chapter 2:
• Exposing incompleteness in long term physical activity data
iStuckWithIt - Study of existing long term trackers and User Interface Design for Long Term Physical Activity Tracker data
Study 2: 21 Existing Long Term Trackers
2. How existing trackers use their own data.
3. iStuckWithit: user interface designs for long term physical activity data.
Chapter 3:
• How existing long term physical activity trackers use their own data
• iStuckWithIt - user interface designs for long term physical activity data
Chapter 4:
• Designing for Adherence: Scaffolding (iStuckWithIt)
Defining Adherence: Measures Of Incompleteness
Study 3: Analysis of 12 datasets containing 753 unique individuals with 77,000 days with data interspersed with 73,000 days without data
4. Exposing the gaps in methods of dealing with incompleteness.
5. Defining adherence, measures of incompleteness.
Chapter 5:
• Exposing the gaps in methods of dealing with incompleteness
• Defining Adherence: Measures of Incompleteness
Chapter 6:
• Implications for personal informatics applications

Figure 1.1: Map of the thesis: Studies conducted, their contributions and the respective chapters that describe these studies and contributions.

However, this thesis takes a different approach. This thesis aims to harness long-term physical activity tracker data even when it is incomplete. This is important because the literature indicates that many people's long-term physical activity data is incomplete [Epstein et al., 2016b, Meyer et al., 2017, Shih et al., 2015].

1.2 Key Contributions & Thesis Structure

This thesis explores how we can help people to make sense of long-term physical activity tracker data. Figure 1.1 shows a map of how this thesis evolved over the three studies conducted. The figure shows how these studies map to the five contributions and chapters in this thesis.

The first study, Study 1 in Figure 1.1, was a preliminary exploration of how people use their trackers over an extended period of time, and the patterns in their physical activity. What really stood out when analysing the data from the 237 students over the semester was just how important the varied wearing behaviour was. It became clear that any reporting on activity levels should account for the uncertainties that occur as a result of different wearing behaviours within a cohort over the long-term.

Study 2 in Figure 1.1 was a three-part experiment involving 21 participants who had been wearing a tracker for at least six months. The design of this second study was influenced by the first study's findings on tracker wearing behaviour and the resulting incompleteness that can occur in activity tracker data. As such, interviews were first used to explore how these existing long-term trackers made use of their long-term data and what they understood about their own long-term physical activity. The second part of this experiment evaluated the *iStuckWithIt* interface. The *iStuckWithIt* design was also influenced by the notion that incompleteness, which is a form of uncertainty, should be exposed and accounted for especially when helping users reflect on their own long-term physical activity data. The third part of this experiment assessed the benefits of a self-reflective scaffolding panel that was intended to help people consider aspects that are important for understanding their physical activity levels. Three contributions resulted from this study: 1) new insights into the data use of people who are long-term trackers; 2) the design of a user interface for self-reflection on long-term physical activity data; 3) evaluation of this data in terms of the insights participants gained from it, both unaided, and with scaffolding.

The first two studies really highlighted the importance of dealing with incompleteness. In Study 3 shown in Figure 1.1, we took a step back and examined how to deal with incompleteness in long-term physical activity data. This involved review of existing work on accounting for data incompleteness due to tracker wearing behaviour. The contribution of this study is based on our definition of *adherence*; this study explains how measures of completeness can be used as a framework for addressing the challenges presented.

Chapters 2 to 6 of this thesis are a series of publications reporting these studies and contributions. In the following sub-sections, a more detailed summary of each chapter is provided, including background, a summary of the key findings and how the chapter contributes towards the thesis goal of supporting people to make sense of long-term physical activity data. We also highlight the key aspects of the reviewed literature in each chapter and its place in this thesis.

1.2.1 Chapter 2: Exposing incompleteness in long-term physical activity data

Early in our research, we found that perceived inaccuracy of data from activity trackers can be a key barrier to continued adoption. The literature reviewed in this chapter introduces the challenge that incompleteness in activity tracker data poses to perceived accuracy, and ultimately, the trust and usefulness of the data collected by trackers [Consolvo et al., 2014, Lazar et al., 2015, Yang et al., 2015, Harrison et al., 2015].

Our review revealed few studies that had analysed data from large cohorts of individuals wearing personal activity trackers over periods of time of more than a few days [Cadmus-Bertram et al., 2015a, Shih et al., 2015], and none of these reported in detail their wearing patterns. Thus, it was unclear how long each individual wore their trackers for during the day, and whether tracker wearing behaviour was consistent among individuals, or across groups and over time. Moreover, there were no analyses of the potential impact of wearing patterns on the interpretation of long-term physical activity data in relation to questions such as the examples in Table 1.1 e.g., do wearing patterns affect the determination of the answers to key questions about a population's average daily step count?

In this first study, we examined the wearing behaviours of two large cohorts of university students during a semester [Tang and Kay, 2016]. A total of 237 students from information technology (IT) and medical science (MED) courses were provided with a loan Fitbit Zip device for the duration of a university semester, which included a mid-semester break. The sample was a convenience samples; nonetheless, the two cohorts were large and were characterised by distinct differences in physical location within the university, as well differences in social environments due to differences in courses studied. These are detailed in Chapter 2.1.2. In analysing the data from these cohorts, we introduced two new completeness measures: *daily and hourly adherence*. Daily adherence refers to the number of days where a user recorded at least one step, and hourly adherence refers to the number of hours in a day that the user recorded at least one step. We hypothesised that these measures would enable us to examine wearing behaviour in more detail and would expose challenges in to making sense of physical activity tracker data.

This study revealed two important challenges to perceived accuracy. First, as expected, even on days where users did wear their tracker, many did not wear it for the whole day: we defined this as wearing the tracker for less than 10 hours. Based on this, 10 hours of data was defined as adherence; IT and MED students differed in the percentage of users who consistently wore their tracker the whole day during the study period. Second, we showed that there were similarities in wearing patterns between IT and MED students (e.g., lower daily adherence on weekends and mid-semester breaks), as well as significant differences between the cohorts (e.g., overall adherence levels).

This study is our first to use the term *adherence* to describe completeness. ***Adherence*** is based on the notion that an activity tracker data should give accurate answers about activity for people who have worn their tracker all day, every day, i.e., 100% adherence. Adherence measures, such as daily or hourly adherence aim to describe the level of completeness in wearing behaviour, i.e., less than 100% adherence. There are many reasons why people may not want to, or be able to, wear their tracker all day every day. For example, a user may be restricted by their work requirements or sportwear requirements, or they may only be interested in tracking data when exercising or tracking during weekdays at work. Adherence is a core underpinning concept in this thesis; it was further explored in Study 2, as shown in Figure 1.1, culminating in a deep exploration in Study 3.

1.2.2 Chapter 3: How existing long-term physical activity trackers use their own data

In our quest to help people make sense of their long-term physical activity data, the paper in this chapter [Tang and Kay, 2017] examined how existing long-term trackers use and make sense of their own long-term data. We interviewed 21 existing physical activity trackers, i.e., people who had used trackers for six months or more (average 23 months; 17 participants averaged one year or more). In addition, we conducted a qualitative analysis of the rich responses of these participants to interview questions about their self-knowledge of their activity levels and actual use of their long-term physical activity data.

The literature reviewed in this chapter focused on previous work addressing incompleteness in physical activity tracker data. This review also situated this study as the first in-depth exploration of the use of long-term physical activity data by existing users, and the first study to design and evaluate an interface depicting adherence information as part of the user interface for self-monitoring and reflection on long-term physical activity data.

Perhaps the most surprising finding from this work was that while our participants accumulated long-term data, it was mainly as a by-product of short term uses. Most participants simply used their tracker to get a daily and weekly view of their activity levels. Only six (29%) participants examined their data over periods longer than one month and only one participant regularly used his long-term data to set goals. In line with other research [Li et al., 2011, Consolvo et al., 2014, Choe et al., 2014, Bentley et al., 2013, Rooksby et al., 2014], we concluded that a key barrier for them to users exploring their long-term data is difficulty in aggregating this data in a meaningful form. Put simply, even if people want to examine their long-term physical activity data, it is currently difficult to do so.

Another surprising insight from this study was that many of our participants, who were long-term users of activity trackers and reported that they reviewed their data each day, did not have very accurate awareness of their own activity levels. This assertion is supported by two findings. First, when we asked users to estimate their own activity levels, there was a 20% difference (on average) between the estimated and actual activity levels in their tracker data. Second, when asked whether they were more active on weekends versus weekdays, of the 12 who felt able to provide an estimate, six (50%) got it wrong. This included one participant who was very active and closely monitored his own physical activity data.

This study contributes important knowledge on how existing trackers used their own data. Together, the findings highlight the need for further development of tools to help people make use of their own long-term physical activity data.

1.2.3 Chapters 3 & 4: *iStuckWithIt* - User interface designs for long-term physical activity data

A key aspect to helping people to make sense of their long-term physical activity data is the design of suitable systems and user interfaces that make that data available in a form that enables users to answer important questions, such as those in Table 1.1. This paper reviewed personal informatics literature on the landscape of the user interface designs of physical activity trackers, providing a summary of what people wanted to see and the different design approaches, and giving an in-depth review of existing user interfaces for long-term data. This paper provides the background and the motivation for the design of *iStuckWithIt*.

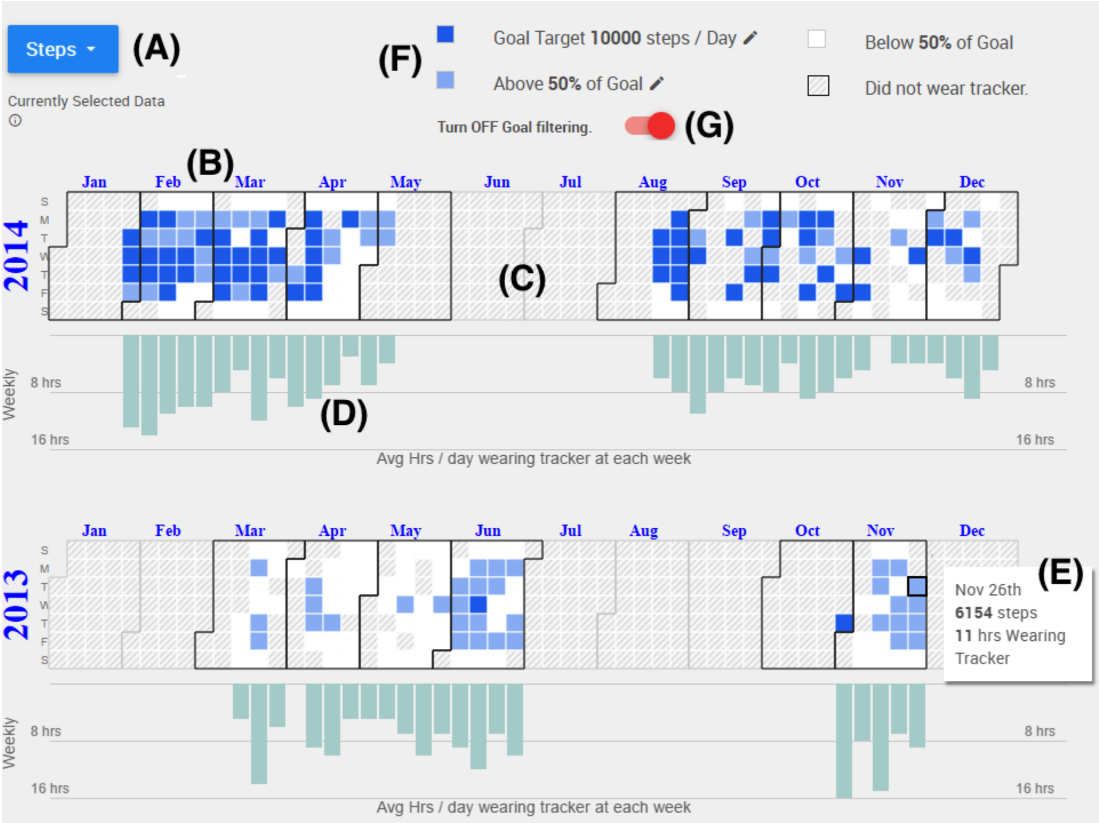
This study aimed to extend the limited available work on user interfaces for long-term physical activity data. In addition, we hypothesised that a suitably designed user interface can be effective in supporting reflection when it also exposes the incompleteness in data (in the form of adherence measures) and relates this to the user's activity levels and goals.

To this end, we developed *iStuckWithIt*, a user interface that allows users to upload then view their own long-term physical activity data. Figure 1.2 [2] shows an overview of the user interface. The centrepiece of this design is the calendar chart visualisation which encodes whether the user reached their goal (dark blue cells), 50% of their goal (light blue) or less than 50% of their goal (white). In this design, we visualised two adherence measures (daily and hourly adherence). Daily adherence is shown by the cell colour: days where the user did not record data are encoded as grey cells in the calendar. The aim of this interface is to provide an overview of actual step counts together with information about incompleteness for long-term data. Hourly adherence is the average number of hours of tracker wear per day, shown as a weekly bar graph below the calendar. The aim of this is to show incompleteness at the daily level. Users are also able to hover over individual cells to view detailed information about their activity level and hours of wear on a particular day. In addition to steps, users can switch the viewed data to active minutes or distance travelled since these are also important measures for reflection over the long-term. Users can also to switch between the goal view (three colours) and the gradient view, where the colour (darkness) of a cell is mapped to a gradient determined by an activity value between 0 (white) and the global maximum (darkest blue), as shown in Figure 1.3 [3]. This gradient view is useful for users who do not have a target goal; this mode was intended to help these users explore their data and then set a goal.

The design of *iStuckWithIt* evolved based on the experiences of our research group in designing of user interfaces for visualising long-term physical activity data [Barua, 2016]. Inspired by works in information visualisation and interactive techniques used within our group, we chose the calendar visualisation as the centrepiece of our design, with interactive techniques to support the exploration of details within the data.

[2] Figure 1.2 also appears as Figure 3.1.

[3] Figure 1.3 also appears as Figure 3.2.



Goal Target **10000** steps / Day

Above **50%** of Goal

Below **50%** of Goal

Did not wear tracker.

Turn OFF Goal filtering.

Nov 26th
6154 steps
11 hrs Wearing Tracker

Figure 1.2: *iStuckWithIt* for hypothetical user Alex for 2013 to 2014. This screenshot shows his long-term steps data against a step goal (A) of 10,000 steps / day (F). (A) Button selects type of data (step count is selected). (B) Calendar heat-map, with colour intensity showing daily goal adherence: dark blue (>=10k steps), light blue (5-10K stops), white (<5K steps, grey for no data). (C) Period with no data. Note: the combination of missing data (grey cells) and goal adherence (3 colours) conveys daily adherence (or when they wore their tracker). (D) Hourly adherence is shown by the weekly bar graph (average hours / day wearing tracker). (E) Mouse-over Tool-tip for detail of a day (i.e., 6154 steps on Nov 26th 2013). (F) Coding key and settings to change goal target. (G) Toggle to switch view which is shown in Figure 3.2. Notable features: low goal achievement on weekends; stopped wearing tracker for three multi-months blocks. 2014 has higher daily and goal adherence than 2013.

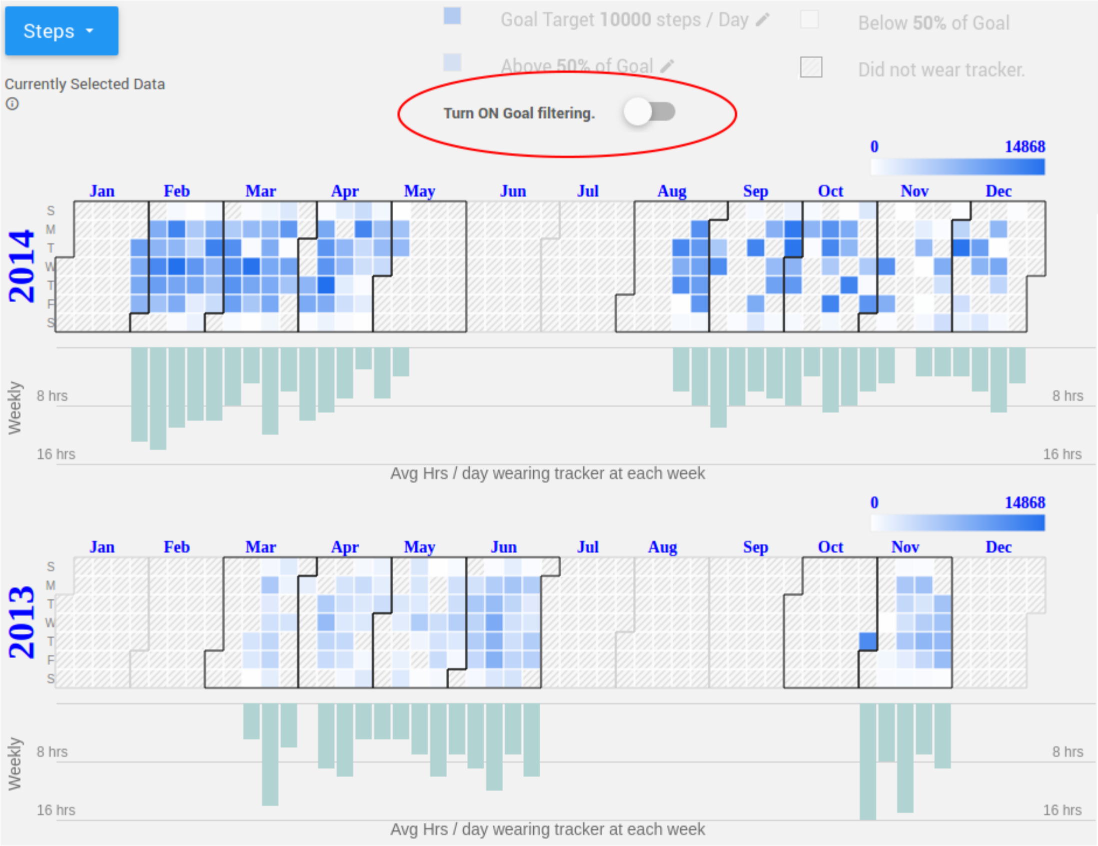


Figure 1.3: ***Gradient*** view of step count (i.e., using colour gradient to denote values from 0 to max step count of 14,868 (upper right chart - colour range legend). This view is activated by the toggle shown in Figure 1.2 (G).

We evaluated *iStuckWithIt* in a laboratory study with 21 existing long-term physical activity trackers users [Tang and Kay, 2017]. We asked the participants to view their own data using a think-aloud protocol while we observed their interactions. Then, we interviewed them about what they had learned and their experiences.

As part of this work, we also explored whether scaffolding support can enhance the core *iStuckWithIt* interface designs [Tang and Kay, 2018a]. We hypothesised that people may need help when making sense of their own long-term physical activity data especially if encountering it for the first time. To explore this, we evaluated two types of scaffolding: goal prompt and tutorial. The goal prompt scaffolding is a side pop-up panel that asks the user five questions about their goals and behaviour, including: are they achieving their goal, should they change goals, with consideration for whether they are at work or and not at work, and if it is a weekend or weekday or a holiday. The tutorial scaffolding asks the user to review the data from two hypothetical users, with mocked data that highlights key concepts within the *iStuckWithIt* design. The scaffolding designs were influenced by literature on metacognitive scaffolding and self-regulated learning technologies [Azevedo et al., 2010, a.W.M.M. Aleven and Koedinger, 2002]. While the premise underlying introduction of the scaffolding was focused on helping students with self-regulated learning, we believe that the scaffolding could be useful in the context of helping people make use of their own long-term physical activity data.

Our work revealed several key insights that can inform the future design of long-term physical activity user interfaces that aim to support self-monitoring and reflection. First, the design of *iStuckWithIt* was very effective in helping users reflect on their own data. Encoding long-term physical activity data in a calendar chart visualisation enabled users to not only view an overall picture, including long-term trends, but also view the detail in their data.

Second, exposing adherence measures (daily and hourly) helped users to reflect on their own wearing behaviour. Many discussed the context and circumstances of non-wear or gaps in their tracking. Moreover, many were surprised or were unaware of the level of incompleteness in their own data. The display of daily adherence was readily understood and valued. The hourly adherence information was less effective. For mnay participants with very high hourly adherence, this pattern of behaviour was constant throughout the study and thus, was not very interesting. Nonetheless, even those with less consistent hourly adherence had difficulty seeing the value of this data.

Third, our study showed that there is indeed a need for scaffolding support. While most users gained insight from viewing their own data using *iStuckWithIt*, the goal prompt scaffold helped many to consider and discover insights that they missed. This work also highlighted potential directions for creating future interfaces to provide personalised scaffolding. We have made several suggestions for future applications in [Tang and Kay, 2018a].

1.2.4 Chapter 5: Exposing incompleteness and defining adherence

The paper in this chapter [Tang et al., 2018] reported on two aspects of our third and final study which focused on dealing with incompleteness in physical activity tracker data. The next two sub-sections first introduce the work by exposing the gaps in the methodology for dealing with incompleteness. Then, define adherence is defined in terms of measures of completeness with the aim of addressing the challenges exposed.

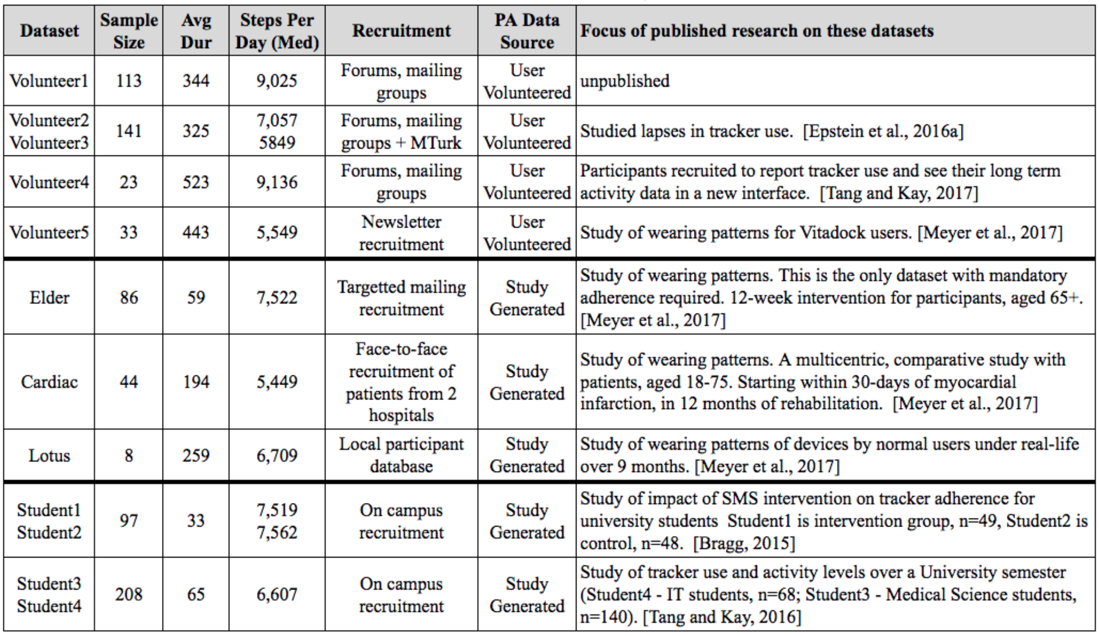
Exposing the gaps in methodology for dealing with incompleteness

Our first two studies led us to ask the question: how appropriate are existing methods for dealing with incompleteness in this type of data? In our third study [Tang et al., 2018], we established a collaboration with leading researchers in the fields of personal informatics and public health. This enabled us to create a rich pool of 12 datasets, containing data from 753 unique individuals; the dataset comprised 77,000 days with data interspersed with 73,000 days without data. This analysis of these datasets makes several contributions. The datasets are listed in Table 1.2 [4].

First, our datasets differed widely in how the physical activity data were collected, ranging from data volunteered by users who had already collected it during their personal use (e.g., Fitbit users) to data from medical intervention studies where participants were asked to wear their tracker during the study period. Second, our datasets varied widely in terms of duration of the study, ranging from 33 days for a student dataset to over 300 days for personal trackers. Finally, the populations represented in our datasets were diverse, ranging from young healthy students to elderly people involved in medical interventions. The wide variation in these datasets allowed us to thoroughly analyse and compare methods for dealing with incompleteness.

[4] Table 1.2 also appear as Table 5.5.

Table 1.2: The 12 datasets from 9 studies of various lengths and population size. The first column is the identifier we use to describe the dataset. Next is the sample size and average duration in days, the average step count (using only days with >0 steps) and then the recruitment methods. The data source column distinguishes *volunteers* datasets (the first block), from the remainder, being *other study-generated* datasets.



Dataset	Sample Size	Avg Dur	Steps Per Day (Med)	Recruitment	PA Data Source	Focus of published research on these datasets
Volunteer1	113	344	9,025	Forums, mailing groups	User Volunteered	unpublished
Volunteer2 Volunteer3	141	325	7,057 5849	Forums, mailing groups + MTurk	User Volunteered	Studied lapses in tracker use. [Epstein et al., 2016a]
Volunteer4	23	523	9,136	Forums, mailing groups	User Volunteered	Participants recruited to report tracker use and see their long term activity data in a new interface. [Tang and Kay, 2017]
Volunteer5	33	443	5,549	Newsletter recruitment	User Volunteered	Study of wearing patterns for Vitadock users. [Meyer et al., 2017]
Elder	86	59	7,522	Targetted mailing recruitment	Study Generated	Study of wearing patterns. This is the only dataset with mandatory adherence required. 12-week intervention for participants, aged 65+. [Meyer et al., 2017]
Cardiac	44	194	5,449	Face-to-face recruitment of patients from 2 hospitals	Study Generated	Study of wearing patterns. A multicentric, comparative study with patients, aged 18-75. Starting within 30-days of myocardial infarction, in 12 months of rehabilitation. [Meyer et al., 2017]
Lotus	8	259	6,709	Local participant database	Study Generated	Study of wearing patterns of devices by normal users under real-life over 9 months. [Meyer et al., 2017]
Student1 Student2	97	33	7,519 7,562	On campus recruitment	Study Generated	Study of impact of SMS intervention on tracker adherence for university students Student1 is intervention group, n=49, Student2 is control, n=48. [Bragg, 2015]
Student3 Student4	208	65	6,607	On campus recruitment	Study Generated	Study of tracker use and activity levels over a University semester (Student4 - IT students, n=68; Student3 - Medical Science students, n=140). [Tang and Kay, 2016]

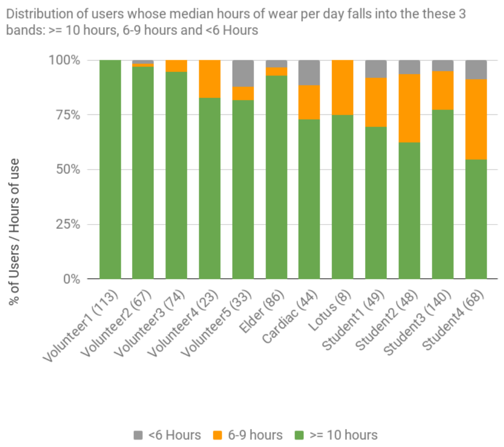


Figure 1.4: Comparison of %-age of users with median wear-time >=10 hours, >=6 and <10 hours and <6 hours. N included in X-axis label. Note: this figure also appears in Chapter 5 as Figure 5.3

A key contribution of this work was our discussion of the limitations of previous approaches for analysing and reporting on physical activity tracker data characterised by wide differences in tracker wearing behaviour.

Our review of the literature revealed that much previous work has been based on datasets where days or people were excluded if they did not meet a certain threshold (e.g., at least 10 hours of steps data, at least 5 out of 7 days a week).

Our analysis demonstrated that simply using thresholds to exclude data can lead to biases, meaning that the results may be more representative of a subset of the sample who are more likely to wear their trackers (e.g., higher adherence to wearing) but are not necessarily more or less active. Figure 1.4 illustrates this challenge, showing the differences across our datasets and the wide variation in the proportions of users who averaged different levels of wear-time (less than 6 hours per day; between 6 and 9 hours a day; and greater than 10 hours per day). For example, only half of the IT students (*students4* on the right) averaged 10 hours of wear time per day compared to 100% of self-motivated Fitbit users (*Volunteer1* on the left).

These variations can significantly impact the results of analyses addressing questions like those in Table 1.1. Figure 1.5 illustrates this. When we use a very lenient, and more inclusive threshold (e.g., days with at least one step), this can result in a significantly different median step count compared to a more restrictive threshold (e.g.,

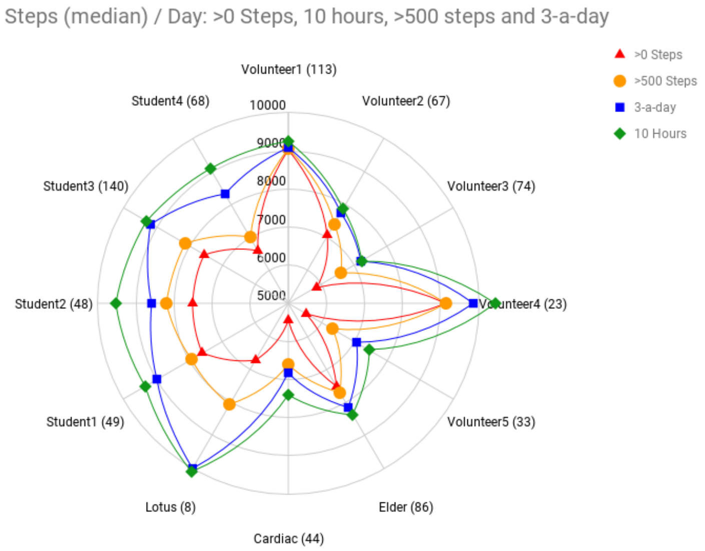


Figure 1.5: Comparison of median steps across populations, showing impact of different valid day thresholds. N included in dataset labels. Note: this figure also appear in Chapter 5 as Figure 5.5

greater than 10 hours of step data). While this difference is negligible for datasets with high tracker wearing adherence to wearing trackers (e.g., Volunteer1 in the centre middle where the difference is so small it cannot be seen in the figure), it can result in important differences for those with lower levels of tracker wearing adherence to wearing trackers (e.g., Student4 in the top left).

This study provides the first consolidated analysis demonstrating the limitations of existing methods for dealing with incompleteness in long-term physical activity data. Given that activity tracker datasets are likely to vary substantially across dimensions impacted by factors such as motivation of use, wearing habits and sample, a new approach is needed if we are to enable people to make sense of this type of data. This work contributes to the analysis of aggregate data in particular, in order to answer questions such as those illustrated in the right-hand side of Table 1.1. However, such data can also be important for individuals, who can make better sense of their own data by comparing it to aggregate data from similar people.

Defining adherence: Measures of completeness

Our work led us to recognise that it is important to address gaps in the methodology for dealing with incompleteness. To do this, we defined ***adherence: a measure of completeness of physical activity tracker data***. Adherence measures provide a well-defined way to characterise completeness in long-term physical activity datasets, analyse this incomplete data, attribute meaning to the results (e.g., hours of wear per day, percent of valid days, percent of users excluded). Instead of using data quality criteria to simply exclude days or data, we proposed the use of adherence measures to analyse the impact of data completeness (e.g., >=10 hours vs >0 steps) and the reporting of analysis results together with adherence measures to communicate the completeness or level of uncertainty within the results reported.

In Chapter 5, we present our definitions of the adherence measures. As part of this contribution, we provide a set of guidelines, with illustrative examples, for using and reporting on adherence measures when working with long-term physical activity data. The guidelines are given at the individual level, for personal health and well-being applications designed to help users achieve their health and well-being goals, as well as for aggregate or population level analysis of long-term tracker data.

Dealing with incompleteness in tracking data is relevant in a number of health research fields including public health and medical intervention studies. Adherence or compliance with wearing and collecting tracking data can vary significantly, and this can have different implications depending on the research participants, goals and methodology used. Incompleteness in tracking data is also important to personal informatics research which collects, presents and helps people make sense of such data. However, the terminology and definitions used in these fields are not common or standardised. Chapter 5, Section 5.2 reviews common terms across health research, including medical intervention studies and public health, as well as personal informatics research. This chapter situates our definitions of adherence in relation to existing practices with examples. Table 5.2 lists common terms used and provides references to key literature with examples, while Table 5.3 lists our definitions.

1.2.5 Chapter 6: Implications for personal informatics applications

This chapter provides a set of guidelines, with examples, on how adherence measures can be incorporated into reports of personal health and well-being applications. First, we recommend that systems and applications should define the adherence measures they have used (e.g., inclusive criteria such as >0 steps versus whole day criteria of $>=10$ hours of use, calculating daily and hourly adherence). They should report the adherence measure and its impact in order to provide insights into the wearing pattern or behaviour of a user or sample and the resulting levels of incompleteness. Then, they should consider how to expose the resulting uncertainty due to data incompleteness to the users. We also provide examples of several ways in which do this, and highlight opportunities to personalise feedback for users.

The discussion presented in Chapter 6 extends work from [Tang et al., 2018]. It provides further discussion on the challenges of data incompleteness and the risks of not accounting for it in personal applications designed to help people with their health and well-being goals. This chapter highlights gaps in existing commercial user interface designs, where these problems are due to a failure to deal with incompleteness, and illustrates the benefits of incorporating adherence measures into these designs (with relevant examples).

1.2.6 Chapter 7: Conclusion and future directions

In this chapter, we conclude with a summary of our contributions, followed by an exploration of the implications of the work of this thesis, a discussion of how these findings fit with emerging technology trends, and recommendations for future work necessary if we are to help people make sense of their own long-term physical activity data.

Chapter 2

Exposing incompleteness

Preamble

> Tang, L. M. and Kay, J. (2016). Daily & hourly adherence : towards understanding activity tracker accuracy. *CHI '16 Extended Abstracts on Human Factors in Computing Systems*

This paper was published in conference extended abstract on Human Factors in computing systems as part of the Computer Human Interactions conference extended abstracts in 2016. It reported on contributions described in section 1.2.1.

Abstract

We tackle the important problem of the accuracy of activity tracker data. To do this, we introduce the notions of *daily* and *hourly adherence*, key aspects of how consistently people wear trackers. We hypotheses that these measures provide a valuable means to address accuracy problems in population level activity tracking data. To test this, we conducted a semester-long study of 237 University students: 88 Information Technology, 149 Medical Science. We illustrate how our adherence measures provide new ways to interpret data and valuable insights that take account of tracker data accuracy. Finally, we discuss broader roles for daily and hourly adherence measures in activity tracker data.

2.1 Introduction

Using activity trackers to improve health is promising [Bravata et al., 2007]. However, to understand patterns of use of trackers, we need to understand the accuracy of the data. One key contribution to inaccuracy follows from the fact that even the most motivated user does not wear their tracker all the time over years, months or weeks. To really understand how active people are, one must account for the fact that inconsistent wear gives an incomplete picture. To address this, we introduce the notion of *adherence* to capture key aspects of the level of wear and show how to use this to give a more accurate picture.

Daily adherence is the percentage of users who wore their trackers each day. Figure 2.1 illustrates this for a population of 237 people over a 2 months period. The figure shows an overall steady drop and also cyclic patterns. Previous literature has reported overall drop-out rates, the rate at which people *stop* wearing their trackers [Endeavour, 2014, Shih et al., 2015]. However, this body of work ignores the accuracy of tracking data during periods when people did wear their tracker.

Hourly adherence is the number of hours users wore their trackers on days they remembered to put them on. It is important to consider this in addition to daily adherence because it reveals how valid the tracking data is on each day. We took the term adherence from its use in medical intervention research [Tudor-Locke et al., 2015] where the data from study participants is only used if they use the device the required number of hours on each of the required number of days. Rather than simply use adherence

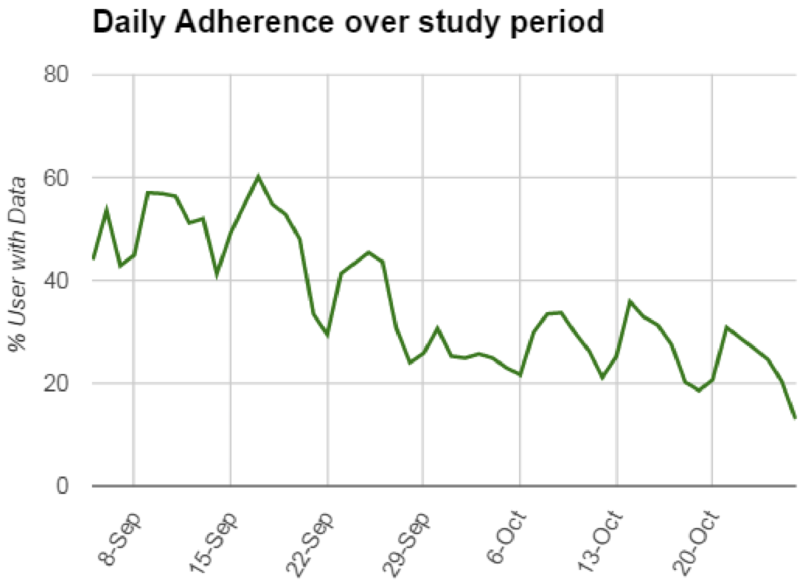


Figure 2.1: Daily adherence (percentage of users with data) over the study period. N=237

as an exclusion criterion, we now show how to use it to make sense of the data that is available from the many people who elect to use trackers as part of their normal lives. This data has the potential to give valuable understanding of populations of users.

While there is a growing body of work on non-medical, general use of activity trackers [Shih et al., 2015, Harrison et al., 2015, Clawson et al., 2015, Fritz et al., 2014] we have found no reports of work that considers daily and hourly adherence to tackle the challenge of data inaccuracy.

We hypothesize that:

- Daily and hourly adherence is important for the accurate interpretation of long-term physical activity tracker data;

To explore the power of these notions, we conducted a semester long study with 237 university students: IT (88) and Medical Science (149). Each was given a Fitbit Zip device and we analysed the data to determine their daily and hourly adherence.

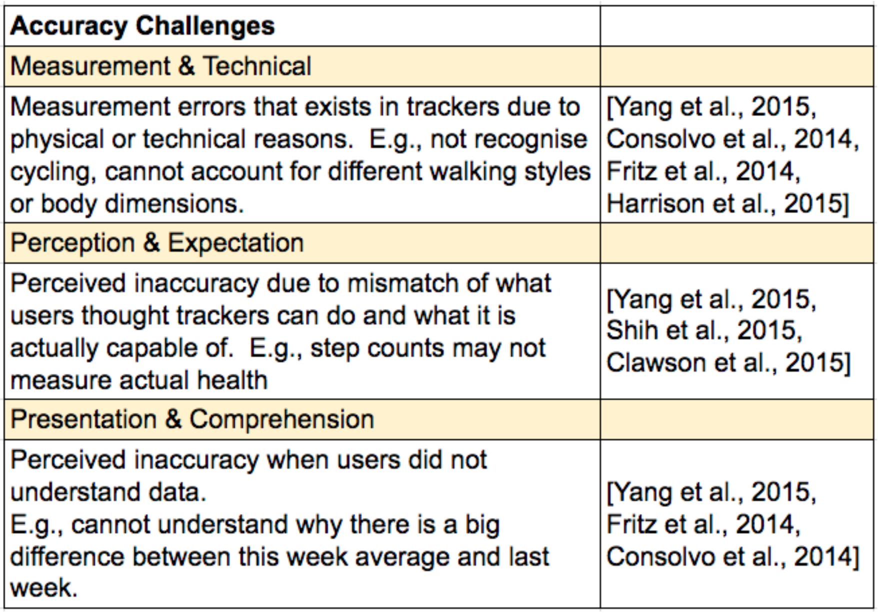
We are the first to report on the daily and hourly adherence levels and patterns on a large group of students. We show how analysis based on daily adherence discloses interesting similarities and differences between the two student groups. This highlights the way that daily adherence has the potential to be a significant source of inaccuracy that can differ across populations. We also show how hourly adherence analysis complements and adds to the picture and needs to be considered as a source of inaccuracy for simplistic analyses of tracker data.

2.1.1 Background

This section first explores the nature of accuracy in the context of physical activity tracking. We then explain the previous use of adherence in such data. Then we show how these aspects link to the main studies of activity tracker use. The section concludes with the positioning of our study to address gaps in the literature.

A number of barriers have been reported in the study of activity tracker adoption including motivation [Shih et al., 2015, Fritz et al., 2014, Consolvo et al., 2014], aesthetics [Shih et al., 2015, Harrison et al., 2015, Clawson et al., 2015], maintenance efforts [Lazar et al., 2015] and accuracy. Our work concerns the last of these. Reflecting its importance, there has been work to understand its forms and impact. We summarise this in Table 2.2, distinguishing three categories of accuracy challenges. Our work tackles the third, presentation and comprehension, which refers to problems due to data being misrepresented or misunderstood. Consolvo used the calorie count presented in many health applications as an example where users are not aware that this value is really an approximation [Consolvo et al., 2014, p. 230]. Moreover, many applications present graphs and summaries which either ignore or do not convey missing data [Consolvo et al., 2014, p. 234]. As we have noted, this is a problem for personal tracker data as it is likely to be incomplete over the long-term [Shih et al., 2015, Consolvo et al., 2014]. Yang et al.reported that users often incorrectly interpreted the inaccuracy in tracking data [Yang et al., 2015]. User daily and hourly adherence data is therefore very important to understand as it directly impacts the accuracy and ultimately presentation and comprehension of the data.

As we have already noted, adherence is used in medical literature on pedometer intervention and health research [Buckworth, 2012, Desharnais et al., 1986, Cadmus-Bertram et al., 2015a]. For example, these work reports using only data where participants wore the device for at least 10 hours a day; any less than this was considered too inaccurate to use. This approach is a good approach to accuracy in measuring intervention outcomes. This is quite different from work on normal use of trackers.



Accuracy Challenges	
Measurement & Technical	
Measurement errors that exists in trackers due to physical or technical reasons. E.g., not recognise cycling, cannot account for different walking styles or body dimensions.	[Yang et al., 2015, Consolvo et al., 2014, Fritz et al., 2014, Harrison et al., 2015]
Perception & Expectation	
Perceived inaccuracy due to mismatch of what users thought trackers can do and what it is actually capable of. E.g., step counts may not measure actual health	[Yang et al., 2015, Shih et al., 2015, Clawson et al., 2015]
Presentation & Comprehension	
Perceived inaccuracy when users did not understand data. E.g., cannot understand why there is a big difference between this week average and last week.	[Yang et al., 2015, Fritz et al., 2014, Consolvo et al., 2014]

Figure 2.2: 3 categories of accuracy challenges reported in recent literature, examples and references.

More generally, the medical literature includes studies of special user populations. Notably, Cadmus-Bertram et al. [Cadmus-Bertram et al., 2015a] reported on a 16-week Fitbit pedometer intervention study with 25 overweight or obese, post-menopausal women. They showed that the median participant wore their trackers for at least 10 hours on 95% of days with no significant decline over time. We have found no reports of such hourly adherence for a broader user population.

Studies of activity trackers use have reported what they call a *drop-off* patterns which describes show long a user wore their trackers before it is abandoned. For example, Shih et al. studied 26 undergraduate students over 6 weeks and found that 65% of participants had dropped off after just 2 weeks [Shih et al., 2015]. Moreover, based on surveys, Endeavour partners reported that more than a third of the owners of smart wearables have abandoned them after 6 months [Endeavour, 2014]. However, the focus on drop-off rates ignores the level of adherence during use and the many ways that people may want to use trackers [Clawson et al., 2015, Harrison et al., 2015, Lazar et al., 2015].

There are good reasons to expect daily and hourly adherence to differ across time. For example, a 7-day study of university students [Sisson et al., 2015] reported lower levels of physical activity on weekends. Daily adherence analysis is needed to account for this. Other studies have shown that various factors affect activity levels, including physical environments [Sisson et al., 2015, Ferreira et al., 2007, Saelens et al., 2012].

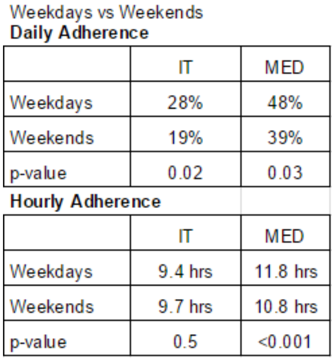
In summary, there is a growing body of work to gain understanding of the ways that people make use of physical activity trackers. In particular, there has been study of drop-off and of the factors affecting both use and drop-off. But, outside medical intervention studies, adherence has not been studied. Yet it seems to have an important role for understanding the ways populations of users actually make use of the devices, giving a more nuanced view than pure drop-off but also pointing to patterns across the week and over long periods of time.

2.1.2 Study Design

To explore our hypothesis, that daily and hourly adherence is important to the accuracy of long-term physical activity tracker data, we designed a study that collected long-term data for two populations of users. Studying daily and hourly adherence across these populations gave us the opportunity to see whether these measures disclosed interesting similarities and differences that impact accuracy. We now describe the populations and the procedures.

We recruited 237 students from 2 university courses: 88 information technology (IT) and 149 medical science (MED). We expected that these students would have different attitudes to activity and tracking. The MED students were second year undergraduates. Their formal studies encourage them to be conscious of health benefits of physical activity. The IT students were third year undergraduates in an HCI subject whose classrooms are at a different part of the university campus. Their studies do not have a health focus. However, the HCI subject had a theme on physical activity, treated in a lecture, homework reading [Church and Blair, 2009, Haskell et al., 2007] and their main assignment was to design a user interface to promote physical activity and reduce inactivity. These groups allowed us to observe adherence differences from students in different social and physical environments.

The Fitbit Zip was provided, on loan for the semester, to each student for the duration of the study. We chose these because they were low cost and had up to 6 months of battery life avoiding maintenance barriers reported in other work [Lazar et al., 2015]. Per minute steps data was obtained through the Fitbit Rest API which allowed us to



Weekdays vs Weekends

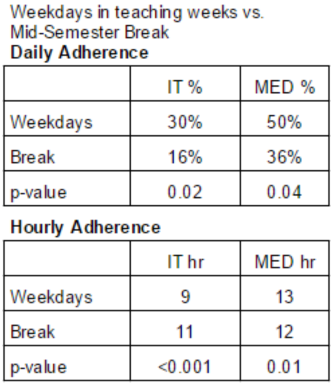
Daily Adherence

	IT	MED
Weekdays	28%	48%
Weekends	19%	39%
p-value	0.02	0.03

Hourly Adherence

	IT	MED
Weekdays	9.4 hrs	11.8 hrs
Weekends	9.7 hrs	10.8 hrs
p-value	0.5	<0.001

Figure 2.3: Mean daily and hourly adherence: weekdays versus weekends.



Weekdays in teaching weeks vs. Mid-Semester Break

Daily Adherence

	IT %	MED %
Weekdays	30%	50%
Break	16%	36%
p-value	0.02	0.04

Hourly Adherence

	IT hr	MED hr
Weekdays	9	13
Break	11	12
p-value	<0.001	0.01

Figure 2.4: Mean daily and hourly adherence: weekdays in the teaching weeks versus the mid-semester break.

determine detailed adherence patterns.

2.1.3 Results

In this section, we report our analysis of the adherence patterns and how these give insights about accuracy of the data sets. We first present and discuss population level daily adherence patterns and discuss our analysis of distinctive features. We then present hourly adherence levels and show how they extend the picture emerging from daily adherence. Finally, we present drop-out patterns, using our data to replicate [Shih et al., 2015] and highlighting how our adherence measures give important new insights into accuracy.

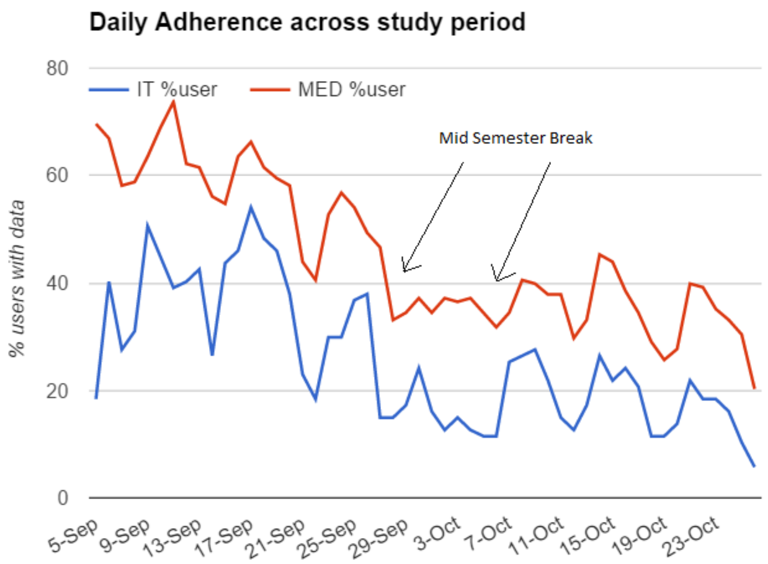


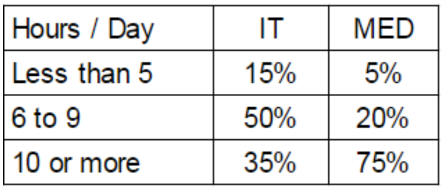
Figure 2.5: Daily adherence (percentage of users with data on a day): IT (blue) vs. MED (red). Note: mid-semester break (30-Sep to 4-Oct). N=237, IT=88, MED=149.

Daily Adherence

Figure 2.5 presents daily adherence of IT (the lower blue line) versus MED students (the upper red line) over the study period. The figure shows weekly cycles through the teaching weeks of the semester. We have labelled semester break; this is flatter than other weeks. Table 2.3 compares means of both adherence measures for weekdays with weekends over the full study. Table 2.4 does this for weekdays in the teaching weeks, compared with the mid-semester break.

Figure 2.5 is a striking demonstration of the way that our daily adherence measure highlights weekly patterns of peaks and troughs. This is remarkably consistent across both student groups. This pattern led us to compare the overall mean daily adherence on weekdays and weekends. This is summarised in the upper part of Table 2.3. For both student groups, there was lower daily adherence on weekends. For IT students this was 19% on weekends versus 28% (p=0.02) on weekdays. A similar pattern, albeit at a high level of daily adherence applied for the MED students, who had 39% on weekends and 48% on weekdays (p=0.03).

Similarly, Figure 2.5's flat section at the mid-semester break (the 1 week (30-Sep to 4-Oct) motivated scrutiny of that week, compared to weekdays in teaching weeks. This shows that daily adherence levels on weekdays of the mid-semester are rather like



Hours / Day	IT	MED
Less than 5	15%	5%
6 to 9	50%	20%
10 or more	35%	75%

Figure 2.6: Percentage of days where the median participant had the hourly adherence levels: less than 5, between 6 and 9 and 10 hours or more.

weekends in the teaching semester. The upper part of Table 2.3 shows that the weekdays in the teaching semester have far higher daily adherence than the break weekdays. Comparing the daily adherence in Tables 2.3 and 2.4 we found no significant difference between the means for all weekends and the mid-semester break weekdays (IT 19% vs. 16% p>0.05, MED 36% vs. 39% p>0.05).

These results support our hypothesis for daily adherence and indicate that accuracy of activity data must take account of the times, such as weekends and also our semester breaks. These results on daily adherence extend the previous findings on lower *activity levels* over weekends [Behrens and Dinger, 2005] to involve lower daily adherence levels as well.

Hourly Adherence

Figure 2.7 shows the IT (lower blue) and MED (upper red) student hourly adherence rates across the study period. This data allowed us to make 2 observations relating to accuracy.

First, many students failed to reach the threshold (10 hours of wear or more) considered valid for medical intervention studies [Behrens and Dinger, 2005, Cadmus-Bertram et al., 2015a]. Table 2.6 reports the percentage of days where the median participant reached 10 hours or more. MED students only reached this level on 75% of days and for IT students it is only 35%. Also, while IT students were less adherent than MED students, they did show some consistency, with 6 hours or more on 85% of days shown in the table. The population level hourly adherence data in Figure 2.7 also shows this quite consistent adherence above 6 hours. We note that the standard deviation levels are high, between 4 and 5 hours, reflecting the large variation within student groups.

Second, IT students wore their trackers for fewer hours than MED students, with

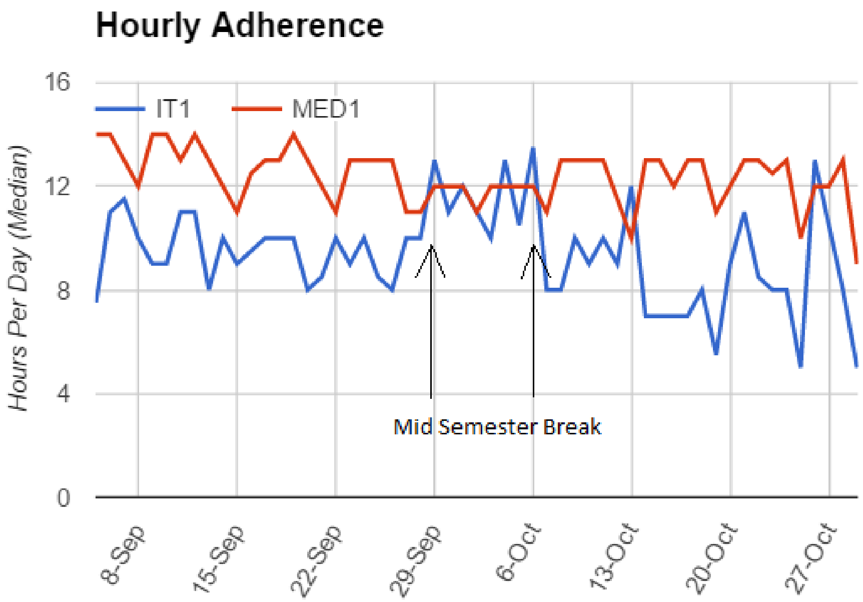


Figure 2.7: Hourly adherence (hours of wear per day): IT (Blue) vs. MED (Red). Note: wide but consistent standard deviation between 4-5hr (not shown). N=237, IT=88, MED=149

means of 9.5 hours and 11.5 hours per day respectively (p<0.001). Table 2.3 shows that MED students wore their trackers for longer on weekdays than weekends. This was not the case for IT students who had similar means of 9.4 hours on weekdays and 9.7 hours on weekends. Combining hourly and daily adherence, we see that while daily adherence is lower on weekends, this was not so for IT students for hourly adherence.

These results support our hypothesis for the importance of hourly adherence in the accuracy of activity tracker data. Moreover, we cannot adopt criteria used by medical research. That is not necessary for personal tracking, and adherence measures can still make population data useful. This makes it feasible to draw on the large amounts of data from many users whose data can still provide valuable insights on real uses of activity trackers.

Drop-off Rate

Drop-off rate is a cumulative measure and refers to the percent of users who completely stopped using their tracker at different times during the study period. We replicated the drop-off analysis in [Shih et al., 2015], with the results shown in Figure 2.8. Our data gives lower rates of drop-off over time than was reported for the 26 students studied

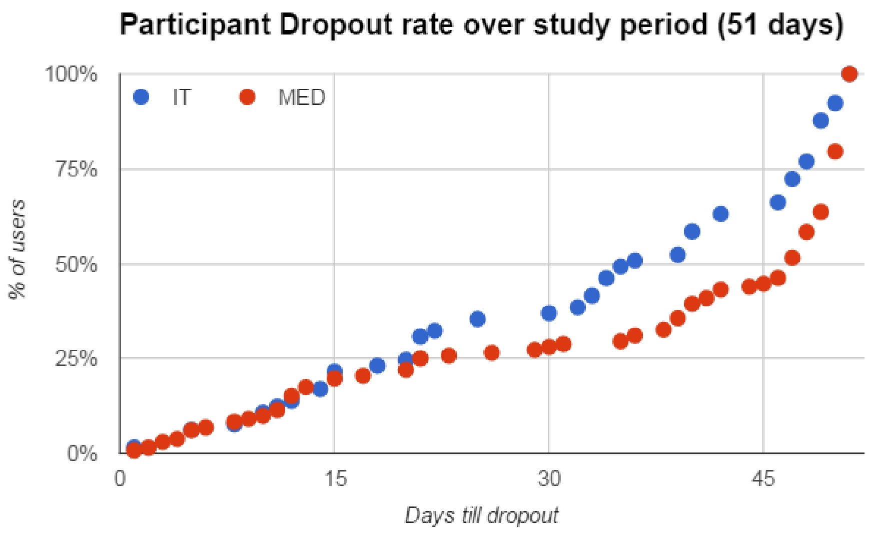


Figure 2.8: Participant drop-out rate of IT (blue) and MED (red) students over the study period. N=237

in [Shih et al., 2015]. This may be due to a number of factors including intervention effects, the populations and environment. It points to the need for further studies for different populations. Studying drop-off rates only indicates when students stopped wearing their trackers. However, it completely omits the effects we have described above.

Limitations

Our study covered 51 days, and different effects may emerge in the longer term. While we had a large population compared to many studies, it is distinctive and can be best seen as adding to the understanding of tracker data. Also, our Fitbit Zip device has limitations, such as limited waterproofing and how it can be worn (i.e., clip on). Notably, the Fitbits were on loan only for the study and this is likely to have impacted results compared with other populations such as those who bought their own devices.

2.1.4 Conclusion & Future Work

We conclude that daily and hourly adherence measures are important. Daily adherence varied significantly at different times such as weekends and the mid-semester break and this impacts accuracy of the population data. Hourly adherence proved to be a potential source of inaccuracy as many failed to wear their tracker for extended hours. Our results point to a need for further work on the perceived inaccuracy at the individual level. Combining population and individual level insights has the potential to offer a clearer picture of adherence patterns relating to a person's own data accuracy. They can help us better determine how and when to apply interventions and also tailoring applications to individual patterns. Accounting for adherence also has the potential to inform design of better interfaces for long-term activity tracker data. Notably, they can take account of daily and hourly adherence in visualisations of longer term data to help them appreciate what their data represent.

2.1.5 Acknowledgements

Special thanks Dr Tina Hinton for her support in working with students on the project.

Chapter 3

Designing for adherence

Preamble

> Tang, L. M. and Kay, J. (2017). Harnessing Long Term Physical Activity Data—How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):26

This work was published in the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies journal in 2017. It reported on contributions described in sections 1.2.2 and 1.2.3.

Abstract

Increasingly, people are amassing long-term physical activitydata which could play an important role for reflection. However, it is not clear if and how existing trackers use their long-term data and incomplete data is a potential challenge. We introduced the notion of *adherence* to design *iStuckWithIt*, a custom calendar display that integrates and embeds *daily adherence* (days with data and days without), *hourly adherence* (hours of wear each day) and *goal adherence* (days people achieved their activity goals). Our study of 21 long-term FitBit users (average: 23 months, 17 over 1 year) began with an interview about their use and knowledge of long-term physical activitydata followed by a think-aloud use of *iStuckWithIt* and a post-interview. Our participants gained new insights about their wearing patterns and they could then use this to overcome problems of missing data, to gain insights about their physical activity and goal achievement. This work makes two main contributions: new understanding of the ways that long-term trackers *have used and understand their data*; the *design and evaluation* of *iStuckWithIt* demonstrating that people can gain new insights through designs that embed daily, hourly adherence data with goal adherence.

3.1 Introduction

More and more people are adopting devices that can track their physical activity. These devices include dedicated trackers, such as FitBit, but they also include multi-function devices such mobile phones and smart watches. Already the first FitBit users could have data spanning 7 years. With time, many people will have long-term collections of physical activity data. The daily data from these devices can help people monitor that day's activity and this may act as a trigger to do more activity. But long-term data has the potential to play other important roles. This is because health improvement and maintenance is a long-term concern. Indeed, any single day's activity is not critical. Nor any one week. It is long-term activity that matters because good health requires lifelong physical activity [Haskell et al., 2007].

long-term data has the potential to enable people to *self-reflect* on their activity levels achieved over the long-term, exploring patterns and features to gain insights into the factors that may have impacted their behaviour [Bandura, 2005b]. This can serve as a foundation for goal setting and planning. Previous work has highlighted the need for better interfaces to support reflection on physical activity data [Choe et al., 2014, Li et al., 2011, Consolvo et al., 2014]. Studies have shown that some people value long-term data [Elsden et al., 2015, Barua et al., 2013], even when they do not yet have a use for it. Various researchers have studied what people would like to learn from their data. They want it to enable them to gain awareness of their actual activity level and how that matches their goals [Epstein et al., 2016a, Choe et al., 2014, Epstein et al., 2014, Epstein et al., 2016b, Consolvo et al., 2014, Gouveia et al., 2015, Harrison et al., 2015, Lazar et al., 2015, Li et al., 2011], taking account of context [Li et al., 2011, Choe et al., 2014] and long-term historical trends and patterns both broadly [Choe et al., 2014, Epstein, 2015, Epstein et al., 2016b, Li et al., 2012], and more specifically, such as to see differences between weekends and weekdays [Behrens and Dinger, 2005, Meyer et al., 2016b, Bentley et al., 2013, Keating et al., 2015].

One key challenge for interpreting long-term physical activity relates to *incomplete data*, due to lapses in wearing the devices all day, every day. Consolvo et al described this as "stuff" happens [Consolvo et al., 2014], Epstein et al referred to this as lapses [Epstein et al., 2016b] and Bentley characterised it as the problem of sparse data [Bentley et al., 2013]. Missing data compromises the meaningfulness of the tracking data. People can lose confidence when they are confronted with gaps and incorrect reports due to gaps [Rapp and Cena, 2016, Bentley et al., 2013]. Indeed failure to account for, or recognise, incomplete data can mean that people consider the data is too inaccurate to be useful and this has been extensively reported in recent years [Rapp and Cena, 2016, Lazar et al., 2015, Shih et al., 2015, Fritz et al., 2014, Yang et al., 2015, Harrison et al., 2015, Elsden et al., 2015]. Epstein et al [Epstein et al., 2016a] reported that 5.9% of their survey respondents abandoned their tracker due to data quality concerns. This is a problem that needs to be addressed in designing interfaces that help people get value from their long-term physical activity data.

To tackle the challenge of incomplete data, we defined *three measures of adherence* to underpin the design of *iStuckWithIt*, a calendar-based visualization of a person's long-term activity levels. Adherence captures the idea of measuring how well people actually adhere to their goal level of activity and use their tracker so that it measures this as accurately as its design permits. ***Daily adherence*** measures how many days a person wears their tracker; we count a day as adherent if the user had any activity data that day. ***Hourly adherence*** is a measure of how much the user wore the tracker each day; we calculate this as the number of hours with at least 1 step within that hour, similar to the calculations done in [Tang and Kay, 2016, Meyer et al., 2016b, Epstein et al., 2016a]. In public health research, it is common to use data only for days with at least 10 hours of data [Cadmus-Bertram et al., 2015a, Buckworth, 2012]. In our terminology, this would be just the days with 10-hour adherence. Daily and hourly adherence provide a way to describe both wearing behaviour and data completeness. For example, a person may have one period of wearing their tracker for most of their waking hours, making it a reliable measure of activity. In another period, where they only wore it for a few hours a day, they may consider the activity levels unreliable. Our third adherence measure, ***Goal adherence***, is defined as the level of physical activity in a day, compared with a target activity level. This concept has been previously described with terms such as goal achievement, step performance. Some people will have their own targets and so will want to judge goal adherence against these. Others may follow a default such as 10,000 steps per day or public health recommendations of 30 minutes moderate activity per day [Haskell et al., 2007]. People may alter their goals over time and want to revisit their data, judging it against a different target. To our knowledge, this is the first work to create an interface that makes use of daily and hourly adherence as a measure of wearing behaviour, linking this to goal adherence. Our work is the first to explore how the lens of adherence can underpin the design of an interface to help people gain insights from their long-term physical activity data, enabling the user to take account of incomplete data.

This paper reports a study of 21 long-term physical activity FitBit trackers (average 23 months, max 38, 17 greater than 1 year). We interviewed them on their tracking behaviour and previous use of long-term physical activity data, to learn about their understanding of their activity levels (average steps) and long-term patterns (weekend versus weekdays). Notably, even in our sample of long-term trackers, 76% had accumulated long-term physical activity data merely as a by-product of daily tracking, not for its long-term value. We compared our participants' perceptions of their activity, from the interview, with their tracker data, which they saw in the subsequent think-aloud with *iStuckWithIt*. Somewhat surprisingly, the participants with highest daily and hourly adherence (people who saw very consistent activity values, day after day, most days) recalled their activity level at a similar accuracy to the participants with lowest daily and hourly adherence. On average, participants mean step estimates had a 20% error (sd: 18%) and half had incorrect understanding of their weekend versus weekday activity levels. Our *iStuckWithIt* interface enabled people to gain new insights about their physical activity, in terms of their patterns of device wearing and activity levels, as well as the context of outliers, patterns and changes in behaviour. Our study demonstrated that people could discover new insights from their long-term physical activityeven when data was very incomplete. Our work makes two main contributions. First, our interview study, before our participants saw *iStuckWithIt*, is the first to report how long-term trackers have made use of their long-term data and their awareness of their physical activity levels. Second, our think-aloud study of *iStuckWithIt* demonstrated its effectiveness for people to gain insights about their physical activity, based on a custom calendar chart, showing *daily, hourly* and *goal adherence*.

The next section reviews previous work. Then we present the user view and design of *iStuckWithIt*, followed by the study design, results and discussion. We conclude with lessons learnt.

3.2 Background

Our goal is to explore how to design a user interface that enables people who have long-term physical activitydata to harness the potential of that data for self-reflection. We first review literature on what people may want to learn from their long-term physical activitydata. We then consider approaches to present physical activity data broadly and their benefits and limitations for the case of long-term data. Finally, we review existing literature on tracker wearing patterns and implications for designing a user interface for long-term physical activitydata. Since our focus is on interfaces onto long-term data, we only include selected aspects of the body of work covering the challenges with adoption and abandonment [Epstein et al., 2016b, Lazar et al., 2015, Shih et al., 2015, Clawson et al., 2015, Harrison et al., 2015] and on the uses of short term data [Fritz et al., 2014, Rooksby et al., 2014, Li et al., 2010, Lazar et al., 2015, Choe et al., 2014, Epstein, 2015]; the key lesson from this literature is that it is important that an interface for long-term physical activityis designed for incomplete data, in terms of both daily and hourly adherence measures.

3.2.1 What people want to see?

We structure this section in terms of the categories of self-reflection identified by Li et al [Li et al., 2011]. Based on their survey of 91 people and more detailed interviews with 15 people who tracked a variety of personal information, including physical activity, they distinguished 6 key classes of questions people asked about their data: status, discrepancies, history, goals, context and factors.

Status and discrepancies. Status refers to gaining awareness of their current activity levels. Discrepancies refers to difference between their current status and their goal. Monitoring status and discrepancies is a common motivation for tracking identified in a number of studies [Epstein et al., 2016a, Choe et al., 2014, Epstein et al., 2014, Epstein et al., 2016b, Consolvo et al., 2014, Gouveia et al., 2015, Harrison et al., 2015, Lazar et al., 2015, Li et al., 2011]. While much of that work concerned the clear value of trackers for short term assessment of status and discrepancies, the long-term form of these is also important for understanding one's activity levels.

History: long-term trends & patterns. Several studies report that some users were interested in their historical data, particularly to find trends and patterns [Choe et al., 2014, Epstein, 2015, Epstein et al., 2016b, Li et al., 2012]. Particularly notable is a study of the current tools, conducted by Elsden et al [Elsden et al., 2015], who invited 15 long-term (> 6 months) personal information trackers of various information to review their own data using their own tools, then describing their experience in an open ended way. Participants often reflected on changes, making meaning from data and reminiscence of moments and periods of life. They suggest that simply helping people *experience* or revisit historical data can be a useful feature for personal informatics tools.

Choe et al [Choe et al., 2014] studied quantified-selfers, those who tracked many kinds of data about themselves and reported on what these users did with their tracker data. They reported that people wanted to see long-term trend and patterns, correlations and relationships between data [Choe et al., 2014].

Li et al [Li et al., 2011] studied a short term interface that had a time based presentation format to present physical activity levels for reflection. A key response from their participants was that they also wanted to see long-term trends, between months, seasons and even years. They suggest that further investigation is needed into how people experience their physical activity data over time.

Goals. As noted by Li et al [Li et al., 2011], tracker data can help people both track goals and identify them. For example, data can help an individual establish their baseline, by reflecting on their previous levels of activity. Then, they can use this to plan for future goal targets, by assessing if their current activity level is a problem and determining what actions they should take to fix this. The importance of goals is also reflected in the work of Rooksby et al [Rooksby et al., 2014] who simply refer to goal driven tracking as a style of tracking where users are motivated by the desire to achieve, or monitor towards a specific target. In some cases, goal tracking motivates short term use of trackers. Lazar et al [Lazar et al., 2015] suggests that in some cases, abandonment can be considered as short term uses because users may only be interested in tracking or learning about their activity data within a specific period of time and not needed when they have reached their short term goal. Rooksby et al [Rooksby et al., 2014] referred to this as diagnostic tracking. Epstein et al [Epstein et al., 2016a] studied why people lapsed (stopped) tracking. In their survey of 141 activity trackers, many participants reported that they abandoned tracking because they thought that they had *learned enough*. This makes good sense as long as people can be confident that their activity levels are stable; however, if they have significant life changes or find a need to become more active, this may not be true and they may need another such period to establish a new baseline. They also highlighted problems with data quality as one reason for abandonment as well as the effort of maintenance. The character of this problem may change as tracking is supported by devices like watches, phones, as well as dedicated trackers that are more convenient to wear. Our work aims to tackle the matter of data quality, especially due to low hourly and daily adherence.

Reflecting on context and factors affecting behaviour. Li et al [Li et al., 2011] refer to context as what other things were happening around the time of their activities or events. They refer to factors affecting behaviour and outcomes over a long period of time. Their participants reported that some depended on memory of events to remember context which they argue is problematic due to the unreliable nature of memory. They also noted that participants found it difficult to explore data to identify factors affecting their behaviour. Li et al [Li et al., 2012] found people were particularly interested in seeking the context around peak performances. Similar results were reported in [Huang, 2016]. They suggest more tools are needed to help people explore their data holistically. Indeed this challenge exists even for many quantified-selfers who are generally considered more advanced in their tracking and tool use [Choe et al., 2014].

Gouveia et al [Gouveia et al., 2015] studied combining contextual data with physical activity data. They reported on a 10-month trial of their mobile app, Habitio, which supports 3 strategies: goal setting, contextualising physical activity with location and daily commutes and textual feedback. It is notable that, they found that people were initially interested in contextual feedback (43%) but this dropped to 18% during the twelfth week of use and overall (38%). They also found that people did not focus on long-term data which seems related to their observation of the ways people used the Habitio app; interaction with contextual feedback was very short, and most related to looking at past day feedback and 71% of participants simply wanted to look at distance walked.

Summary. The previous work highlights the importance of goals, with long-term data potentially helping to set a baseline as well as for revised targets. Partially linked to this, people want to be able to explore long-term trends and patterns and to link these to memories, events and factors related to activity levels. All this needs operate in the context of missing data, to enable people to make sense of their long-term data, even though they may not wear their trackers all day and every day.

3.2.2 Interfaces to present activity data for reflection

There have been many interfaces for short term physical activity data. These range from the early work like Ubifit Garden [Consolvo et al., 2008] and FishN-Step [Lin et al., 2006] to the many interfaces that are available with trackers and associated phone apps. There has also been exploration of informative art by Fan [Fan et al., 2012]. Such short term data interfaces design goals need to support awareness and motivation, rather than long-term self-reflection that is our focus. We now review work that informed our design.

There has been important work to identify *barriers* people have experienced in using existing interfaces to self-reflect. The 2010 work by Li et al [Li et al., 2010] surveyed 68 and interviewed 11 participants to identify several barriers including lack of time, the visualization, perceived criticism, difficulties of interpretation, search, lack of context, sparse and missing data and data that was not useful. In 2014, Choe et al [Choe et al., 2014], studied problems experienced by quantified-selfers. These dedicated trackers most often used a spreadsheet and custom built systems (79%).

Q-selfers reported difficulties in combining data sources and exploring their data for self-reflection. Li et al [Li et al., 2011], in a study of ubicomp technology support for reflection, reported that existing trackers experienced problems in understanding their history, seeing trends and patterns because some lacked long-term data and for others their existing interface made it difficult for them to explore their long-term data. Rapp et al [Rapp and Cena, 2016], in a study of inexperienced personal informatics trackers, found challenges in managing, visualizing, and using their data. They found that the lack of suggestions on using data and the excess of abstract visualization in the apps prevented users from gaining useful insights. These barriers remain for harnessing long-term physical activitydata and we now turn to work towards interfaces to overcome them.

Epstein et al [Epstein et al., 2014] explored *various ways to present activity data*, which they call visual cuts. For example, visual cut 1 is a histogram of number of steps walked each day with a focus on past success. As a foundation for the designs, they surveyed 113 trackers to learn the factors these people believed affected their physical activity. This resulted in 11 diverse factors, the most common being work schedule and weather. They then conducted a month-long trial with 14 participants, finding that different people valued different cuts. Epstein et al [Epstein et al., 2016b] surveyed 141 people to study their interface preferences. They identified 3 groups of use: short-term (<6 months), intermittent (3 or more use periods, separated by 30 days of non-use) and long and consistent use (>5 months with any step data) People with the first two patterns preferred cuts aggregated by hour or day, while the third (long and consistent use) preferred cuts which highlighted their long use. Both these studies indicate the potential benefits of personalisation or customization of interfaces based on visual cuts.

One important and elegant approach is Huang's exploration of incorporating physical activity data into a *personal calendar* [Huang, 2016]. This means that the user can readily draw upon context information in the calendar, both in terms of the events and temporal flow. Users can see activity data within a familiar interface that is part of their normal life and also gain flexibility in data granularity, by switching time scale (day, week, month). She conducted a 9-week field trial with 21 people who had up to 2.5 years of data. While they used the interface just with data for those 9 weeks, the study demonstrated the promise of this approach. This is important for our work as it demonstrates benefits of a calendar interface to address some of the context barriers for reflection.

3.2.3 Adherence & wearing behaviour

We now consider work that informs our understanding of missing data, which we call *hourly and daily adherence*. These measure the wearing behaviour of users and can be used to describe wearing patterns. The term adherence is used in medical literature where it refers broadly to how well people follow a recommended regime, such as taking prescribed medications or doing prescribed activity (e.g., 30 very active minutes 3 times per week [Haskell et al., 2007]). For example, Cadmus-Bertram et al [Cadmus-Bertram et al., 2015a] reported on a 16-week Fitbit pedometer intervention study with 25 over-weight or obese, post-menopausal women. They showed that the median participant wore her tracker for at least 10 hours on 95% of days with no significant decline over time. We have found no reports of such hourly adherence for a broader user population. In the context of public health research about physical activity, [Tudor-Locke et al., 2015] data is used only if participants wore the device the required number of hours (typically 10 hours per day) on each of the required number of days (typically including weekdays and weekends because there are important differences between these). We used adherence to interpret population level activity data, over 1 semester, for 237 university students, where adherence facilitated comparisons of two cohorts of students, one from IT and the other Medical Science [Tang and Kay, 2016]. That work also pointed to the potential value of going beyond population level adherence measures to personal use. In another valuable exploration of daily and hourly adherence at the population level, Meyer et al [Meyer et al., 2016b] studied 34 patients recovering from myocardial infraction, as they used trackers for up to a year. Meyer et al described wearing patterns or daily adherence in terms of duration, density / continuity, streaks and breaks. Using these, they report two key wearing patterns: all patients used trackers on some, but not all days per week; this use pattern was consistent throughout the trial, with no drops in wearing. They also reported hourly adherence (which they call the intra-day wearing pattern) as follows: 88% of days had at least 6 hours or more and 77% had at least 1 step during each of the periods, morning, noon, afternoon and evening. These daily adherence measures are higher and more consistent than our population study [Tang and Kay, 2016]. Both of these and the work on lapses, such as [Epstein et al., 2016b] point to the variability to be expected in individual's wearing behaviour. Notably, the focus of this work was designing visualizations that may encourage people to begin wearing devices after a lapse. This is quite different from our goal to support reflection on long-term physical activity, taking account of incomplete data.

Self-reflection about long-term goal adherence is important for self-regulation in planning and self-monitoring [Bandura, 2005b]. While they did not use the term goal adherence, Epstein et al [Epstein et al., 2016b] displayed goal adherence information as the number of days users achieved different activity levels. We can readily describe previous user interface designs in terms of goal adherence and the diverse ways to present it: aggregated (e.g., average over a period); in terms of a goal-target threshold (e.g., superimpose a threshold line over a steps chart over time); an abstracted form (e.g., ambient displays [Lin et al., 2006, Consolvo et al., 2008]); and as a text summary [Consolvo et al., 2014, Bentley et al., 2013].

Summary

There is a considerable body of work identifying what people believe will support their reflection on long-term physical activity. One key gap in the literature is a study of what long-term trackers understand about their wearing behaviour and levels of physical activity. Our work aims to address the gap by interviewing people who have long-term physical activitydata so that we can learn about their knowledge of their daily, hourly and goal adherence. The literature is beginning to build a richer picture of interface elements for self-reflection on long-term physical activitydata, and of the challenges, particularly in terms of missing data. But it clearly points to the need for new interfaces that can help people harness that data, to find new insights from it. We have used the terms hourly, daily and goal adherence in reviewing the literature above. In the previous work, various terms were used to describe aspects of wearing behaviour, such as lapses duration, density / continuity, streaks and breaks. Similarly, the notion of goal adherence has been variously described, for example, in terms of identifying and tracking goals, goal achievement or performance. Our three adherence measures provide a new way to take account of wearing behaviour when interpreting long-term physical activitydata. We now describe the design of our *iStuckWithIt* which was based on these adherence measures.

3.3 *iStuckWithIt* Design

We built *iStuckWithIt* guided by the following design goals:

DG1 Provide an overview of goal adherence, in terms of steps and active minutes per day.

DG2 Provide an overview of daily adherence.

DG3 Provide an overview of hourly adherence.

DG4 Present data to support reflection on long-term trends, events & their temporal context.

We now describe the user view of *iStuckWithIt* and explain the design rationale. In terms of what people want to see, daily goal adherence and its temporal patterns are the most important (DG1). However, as research points to the importance of missing data in terms of daily adherence we wanted to also make this very clear (DG2). Hourly adherence is important for assessing the accuracy of the data and we wanted users to be able to see this as they explored and reflected on their data (DG3). Together these need to support DG4. We now describe the design of *iStuckWithIt* in terms of the design goals.

Figure 3.1 shows the interface as used in the main study of this paper. (Earlier versions were iteratively refined, with small-scale think-aloud [Nielsen, 1994].) The figure is based on data for a hypothetical user, Alex, who wore a FitBit in 2013 and 2014 and had a goal of 10k steps per day.

At the top left, the blue menu button (A) has *Steps* selected. When clicked, a pop-up menu enables users select from a list of datasets to view. These are daily steps, lightly, fairly and very active minutes, loaded from user's Fitbit account.

The most visible feature of the interface is the *daily and goal adherence* visualisation (DG1 and 2). This is a custom design that embeds the calendar chart element from Google charts [1] and use coloured cells that represent each day to display daily and goal adherence. A calendar chart metaphor allows us to embed adherence measures in the context of a calendar format. Importantly, the calendar format was chosen because it should help people recall relevant contexts and factors that were temporal. Figure 3.1 (B) marks this for early 2014. The colour intensity of each cell shows goal adherence for that day, described at (F). Alex has a goal target of 10,000 steps. *iStuckWithIt*

[1] https://developers.google.com/chart/interactive/docs/gallery/calendar

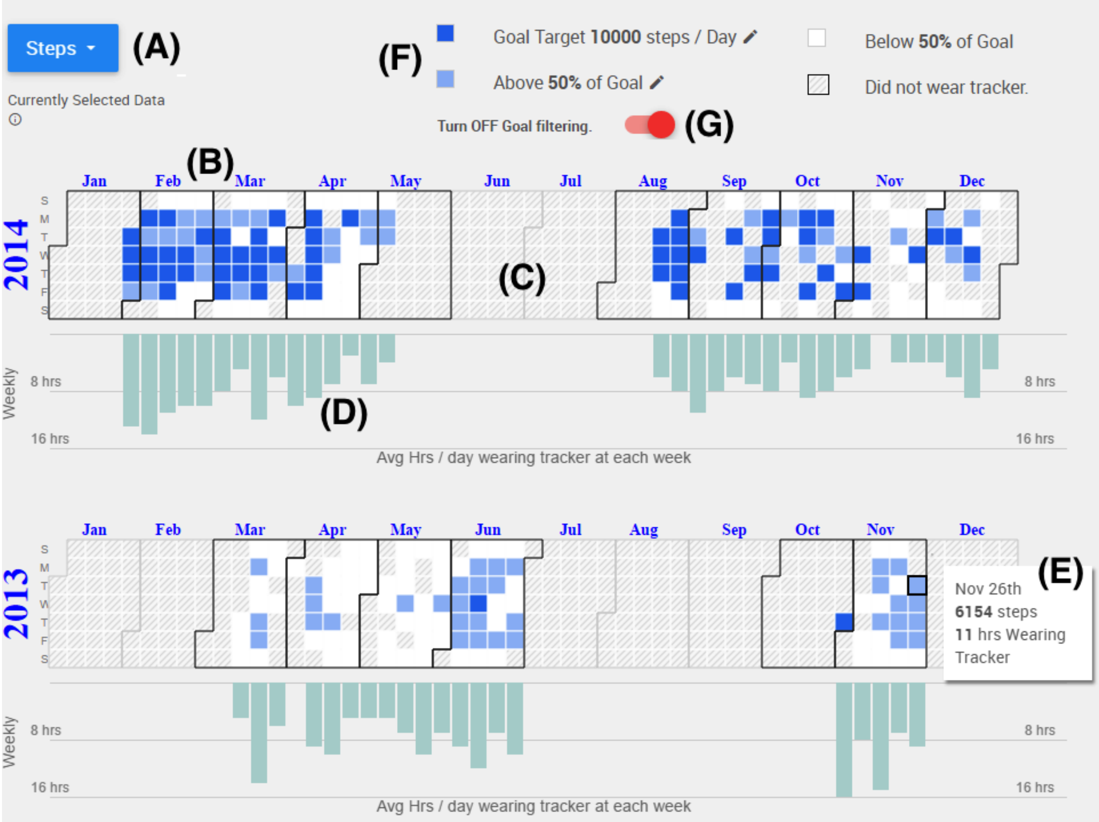
indicates where Alex met the target goal in bright blue. Not meeting it, but achieving 50% of it, is light blue. White indicates Alex had data but his activity level was less that 50% of the target. Grey cells indicate no data. We chose these three levels to make it easier to see differences at a glance. We wanted to clearly show goal adherence days and to be encouraging about the lower level of goal adherence shown in the light blue.

The interface makes it easy to see that Alex had higher goal-adherence in 2014 compared to 2013. He did relatively well in early 2014 (where there are more dark blue cells). We also readily see goal adherence by day, week and longer blocks. For example, in the last week of January, 2014, Alex met his goal on 3 days, with one day at half the target. The daily adherence for that week shows 3 days with no data.

For daily adherence (DG2), we focus on the grey cells. These indicate days the user was not wearing their Fitbit. We designed the interface so the user could readily see both daily wearing adherence and goal adherence so that they could consider their goal adherence, taking account of daily adherence levels and patterns, such as streaks of consistent wear, long breaks, and the various patterns of intermittent days of wear. The calendar shows that Alex started tracking in March 2013 (bottom calendar year), stopping in December 2014 (top) and he had three clear breaks from tracking including 3 months in 2014, shown at (C). Looking across the top and bottom rows for each week, we can see that he had low daily adherence on weekends (grey cells) and even when he did wear the tracker, his step count was below 50% of this goal (white cells).

Under the calendar, the weekly bar chart shows hourly adherence (DG3), as the week's average hours of wear per day (D) in Figure 3.1. This includes data only for those days where there is any data.

The legend at (F) explains the colour coding and the user can click the pen symbol to alter these. For example, a user can alter the goal to 8k steps, and make light blue show days with at least 80% goal adherence. Figure 3.1 (E) shows the *tooltip* where Alex hovers his mouse cursor over 26th of November 2013 to see he recorded 6,154 steps from 11 hours of wear. This aspect of the design provides overview and details-on-demand as recommended for visualization [Shneiderman, 1996]. A *tooltip* is also available when hovering over the weekly average bar graph and shows the hourly adherence values. Making this information only available on tool-tips was a design choice. It follows from our prioritising the overview of goal and step adherence (DG1 and 2) with details of hourly adherence and actual data for each day part of the



Steps ▾

Currently Selected Data

Goal Target **10000** steps / Day

Above **50%** of Goal

Below **50%** of Goal

Did not wear tracker.

Turn OFF Goal filtering.

Nov 26th
6154 steps
11 hrs Wearing Tracker

Figure 3.1: *iStuckWithIt* for hypothetical user Alex for 2013 to 2014. This screenshot shows his long-term steps data against a step goal (A) of 10,000 steps / day (F). (A) Button selects type of data (step count is selected). (B) Calendar heat-map, with colour intensity showing daily goal adherence: dark blue (>=10k steps), light blue (5-10K stops), white (<5K steps, grey for no data). (C) Period with no data. note: the combination of missing data (grey cells) and goal adherence (3 colours) conveys daily adherence (or when they wore their tracker). (D) Hourly adherence is shown by the weekly bar graph(average hours / day wearing tracker). (E) Mouse-over Tool-tip for detail of a day (i.e., 6154 steps on Nov 26th 2013). (F) Coding key and settings to change goal target. (G) Toggle to switch view which is shown in Figure 3.2. Notable features: low goal adherence on weekends; stopped wearing tracker for three multi-months blocks. 2014 has higher daily and goal adherence than 2013.

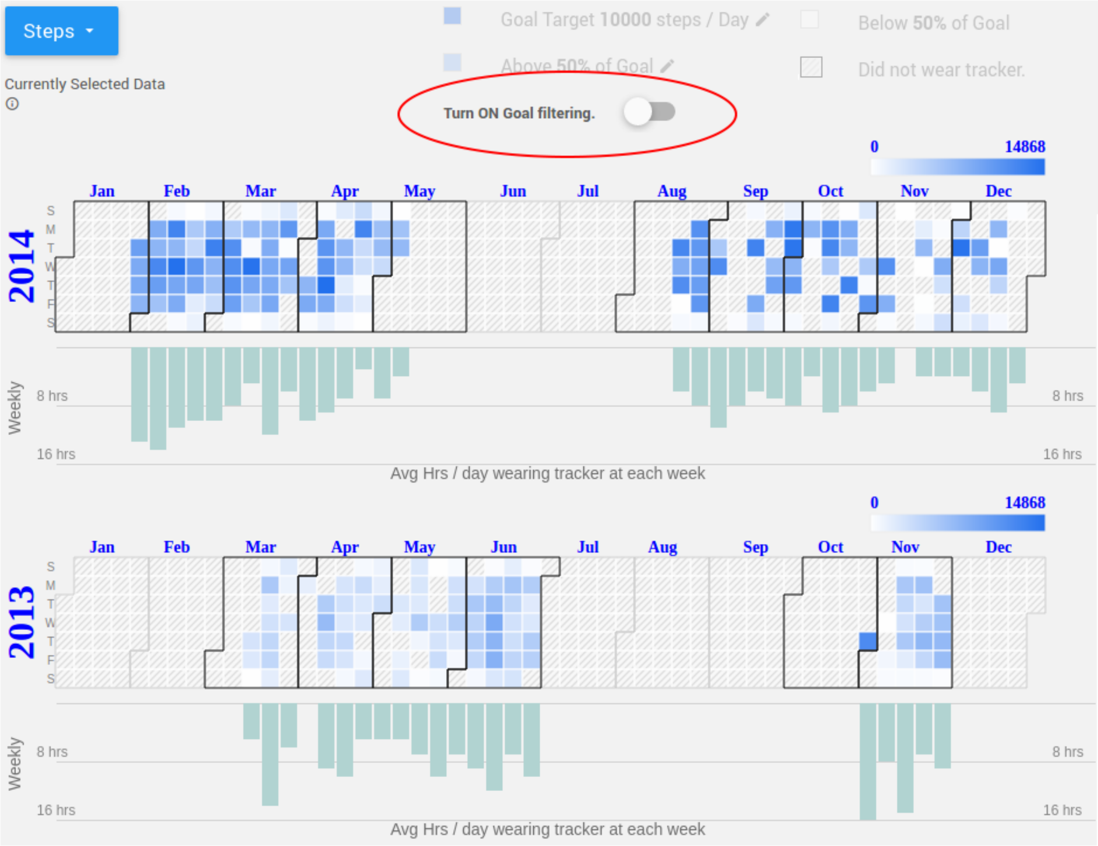
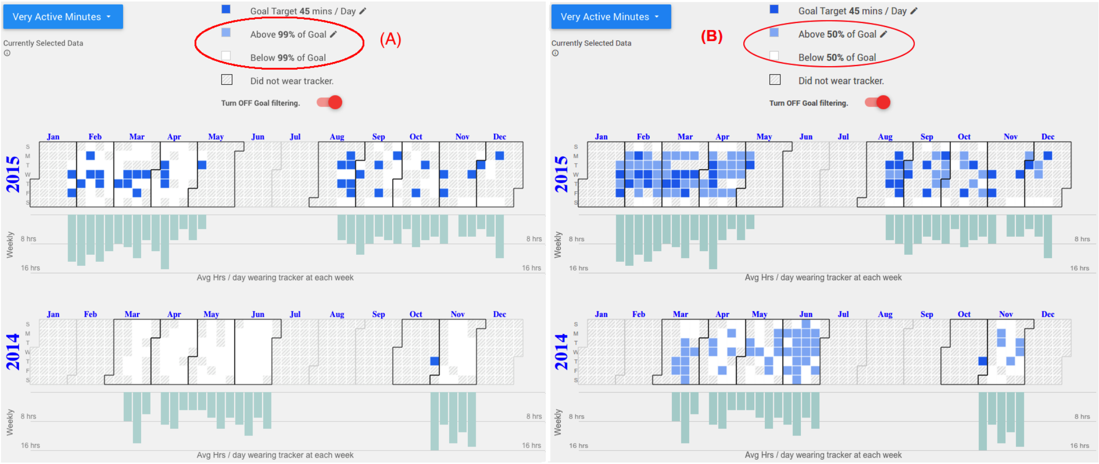


Figure 3.2: ***Gradient*** view of step count (i.e., using colour gradient to denote values from 0 to max step count of 14,868 (upper right chart - colour range legend). This view is activated by the toggle shown in Figure 3.1 (G).

user of their data as they reflect on features that are important to them. So, for example, if Alex is interested in the last week of January 2014 when he restarted wearing his tracker, he can see the hourly adherence is high and the tooltip will reveal that he averaged 15 hours of wear on days with any data.

We now show how the interface design supports flexible exploration of various goal adherence levels. We do this in terms of goals expressed in terms of very active minutes, an alternative to step counts. Figure 3.3 shows a pair of screen shots. On the left, it shows Alex has a goal of 45 minutes of active minutes per day. The settings (A) in the left one makes it easier to focus on full goal adherence (i.e., 45 minutes target, 99% threshold). The right one shows the same user data, now, showing how much light blue appears for 50% goal adherence - see (b) in the figure. These goal settings provide an interactive capability for users to explore their goal adherence with different targets



Goal Target 45 mins / Day

Above 99% of Goal

Below 99% of Goal

Did not wear tracker.

Turn OFF Goal filtering.

Goal Target 45 mins / Day

Above 50% of Goal

Below 50% of Goal

Did not wear tracker.

Turn OFF Goal filtering.

Figure 3.3: Shows very active minutes of Alex with 2 goal target settings (A) and (B). Settings (A) shows how to show only days when users achieved their goal (i.e., 99% of 45 minutes). Settings (B) shows adjusting light blue cell threshold from 99% to 50% of 45 very active minutes goal (i.e., light blue cell for days where very active minutes is between 22.5 and 45 minutes).

and thresholds.

We call the viewing mode shown in Figure 3.1 the *goal filtering* view. This was designed to show goal adherence and was intended for users who have a goal such as Alex's 10,000 steps. We also created an exploratory mode, which we call ***gradient*** view. Previous studies suggested that user interfaces and visualisations should offer the ability to explore personal data not just against specific goals or targets [Epstein et al., 2014, Li et al., 2011]. Clicking on the red toggle in Figure 3.1 (G) shows how users can switch views and a gradient view example is shown in Figure 3.2. This mode transforms cell colours to a gradient determined by activity value between 0 to the global maximum. In Figure 3.2 we can see that Alex's peak step count was 14,868 steps and now the darkest blue cells are close to this while white means step counts close to 0 steps.

Overall, the choice of a calendar format takes advantage of people's existing mental map of calendars. It should also facilitate recollection of relevant events and their context (DG4). For example, in Figure 3.1, Alex had multiple long period lapses during 2013 and 2014. He may be able to recall the circumstances of gaps and reflect on them (e.g., he may have been lost his tracker in July 2013 and only bought a new one in October 2013). Similarly, a user may recognise a period of training for a marathon, a holiday with lots of walking, a change in job that meant he no longer walked to work. Our interface is similar to Huang [Huang, 2016] in its use of a calendar. An important difference is that we do not embed the activity data into a normal calendar. Rather we have carefully designed a calendar chart visualisation to show long-term data to support reflection. It may be useful to link these approaches so the daily visibility of activity data could trigger a user to move to the long-term reflection interface.

The calendar chart and weekly bar graph design choices were guided by the following rationale. First, we chose the calendar chart view because it is an efficient overview + detail on demand visualization that shows many months and years of data together but still is able to show individual days. A calendar format also enables users to see long-term patterns between months, seasons and years. We chose to present hourly adherence in the weekly bar graph. However, our research goal was to examine how users interpret and make use of the weekly hourly adherence patterns in their long-term data.

3.4 Experimental Design

We set out to study how user interface designs that incorporate adherence measures can help people harness their long-term physical activitydata for self-reflection. We designed a study to evaluate the insights people could gain from having a view of their own long-term physical activitydata while using the *iStuckWithIt* application. To do this, we first needed to understand how they currently use their long-term data, what they already believed; this also contributes to understanding of what people know about their tracking and physical activity. We then observed their use of *iStuckWithIt* and report on their insights, experience and design implications.

Our study aimed to answer the following questions:

RQ1 How do people currently use their long-term physical activity data and how does this relate to their actual daily, hourly and goal adherence?

RQ2 What insights can people gain from our interface which shows daily, hourly and goal adherence over the long-term?

We recruited 21 Fitbit users with at least 6 months of data, using social media, internet forums and University mailing lists. Recruitment information outlined the nature of the study, duration of 30-60 minutes and that those completing it be in a draw for a $50 voucher.[2] The research prototype *iStuckWithIt* is a web application which allow users to sign-up to our study and ask them to provide FitBit credentials via OAuth2 protocol which enables our application to download their Fitbit data through the Fitbit Rest Api [3]. The study sign-up and interview steps are detailed below.

Step 1: Background questionnaire and set up.

In this phase, participants registered online and completed a consent form as part of the sign-up process for *iStuckWithIt*. They answered a questionnaire on demographics, current physical activity, exercise stage of change and self efficacy towards exercise [Marcus et al., 1992]. These were to support analysis of the participant's self-knowledge and adherence measures (RQ1). The sign-up process then asks them to link their Fitbit account to our web application *iStuckWithIt* which then initiates the download of Fitbit 1-minute data which can take several hours depending on the amount of data. On completing this phase, those who had been tracking for at least 6 months, were invited for the next steps.

Step 2: Pre-interview.

This was conducted in person or via teleconference, according to participants' location and preference. This pre-interview contributed to RQ1, and was also used to compare participant beliefs about their adherence, activity levels with actual data. It also help us compare previous beliefs with insights from their feedback during the think-aloud and post-interview. We asked about current wearing behaviour, including daily and hourly adherence, whether they believed their wearing and activity patterns differed on weekends and current use of the Fitbit mobile app, website and weekly email report. We also asked them to elaborate on their physical activity and health goals, how important they considered tracking and why they track.

Step 3: Think-aloud session.

In this stage, we asked participants to think-aloud as they used *iStuckWithIt* to explore their own Fitbit. We observed and recorded their comments towards RQ2.

[2] This study was approved by the University of Sydney Ethics Committee, ID-2013/811.

[3] dev.fitbit.com - Note: a special developer authentication token is required to access the 1-minute Fitbit data.

Step 4: Final Post-interview.

This interview asked participants to comment on what they had previously learned from viewing their long-term physical activitydata. We also asked for free comments about the interface, its usability, preferences for viewing mode, understanding of daily, hourly, goal adherence, an assessment of its usefulness and what they liked most and disliked.

Analysis.

To analyse qualitative data, from think-aloud comments and interview responses, we used inductive thematic analysis to identify emerging themes [Dennison et al., 2013]. In all cases where we report statistical significance, $p < .05$.

Limitations

Our study was restricted to FitBit users. As FitBits had been widely available for many years [4], this allowed us to recruit many existing trackers with long-term data. It may well limit generalisation to other devices. Our data does not distinguish between inactivity and non-wear. We treat any days with 0 steps as non-wear and calculate hourly adherence by including any hour that has any steps. This threshold for daily adherence was chosen after analysis of the data indicating that it had similar results to other low thresholds, such as 500 steps. Our approach was also mentioned in previous studies with such devices [Meyer et al., 2016b, Epstein et al., 2016b].

3.5 Results

In this section, we first present profiles of the participants. The next three subsections present analyses related to RQ1, based on the pre-interview and the actual data to reveal their perceived wearing and physical activity behaviour. Then we present results for RQ2, the insights participants gained during the think-aloud use of *iStuckWithIt* and participants' response to the interface. We then report the picture that emerges from their data, for daily, hourly and goal adherence, and how this links to the profiles

[4] http://www.wareable.com/fitbit/fitness-tracker-sales-2015-fitbit-1169

described above. Finally, we report the key insights people gained from using the *iStuckWithIt* application.

3.5.1 Participants

Table 3.1 shows the profiles for our 21 participants, ordered in increasing time since they started tracking (Col. 2). Our sample size and distribution is comparable to similar other qualitative studies of long-term trackers, such as [Fritz et al., 2014] (24 participants $>$ 6 months), [Elsden et al., 2015] (15 participants $>$ 6 months) and [Huang, 2016] (7 participants $>$ 6 months). Seventeen (81%) of our participants had more than 1 year of use with a median of 23 months and maximum of 38 months.

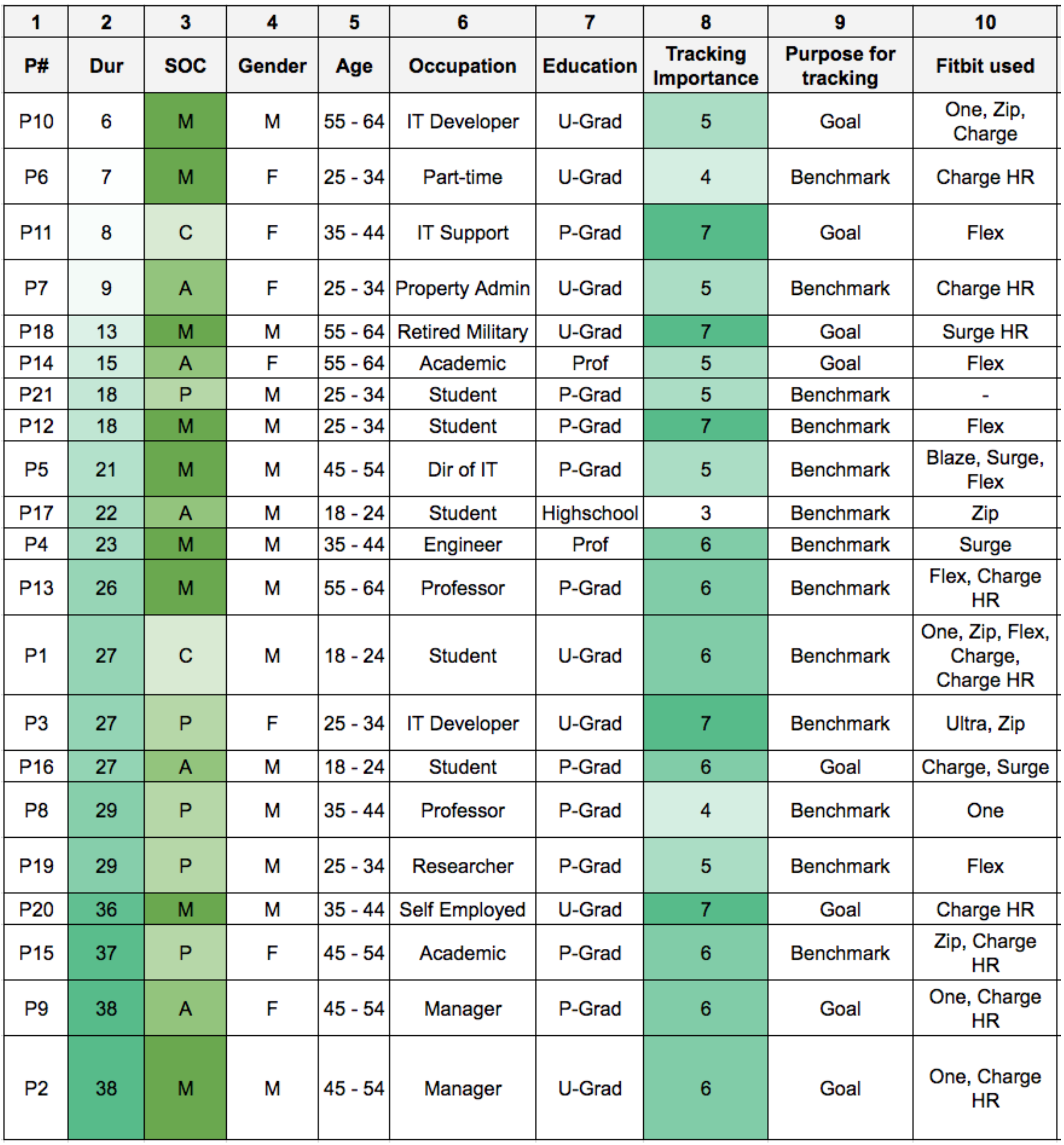
Col. 3 shows the exercise stage of change, ranging from the lightest green for the 2 participants at the contemplation (C) stage, 5 each for preparation (P) and action (A) to the darkest green for the 9 in maintenance (M). It is hardly surprising that long-term trackers have so many people in maintenance phase although these people are spread through the range of tracking duration.

Col. 4 shows that there are 7 women, 33% of the participants. The age groups, in Col. 5, are distributed fairly evenly across each 10-year range from $35-64$, slightly more in $25-34$ and slightly less for $18-24$. While we have more men than women, the age demographic is in line with a survey of 5000 US consumers on fitness tracker use[5]. Col. 6 shows a predomination of IT, research, students and academics, although 6 are from broader occupations and this is a highly educated group (Col. 7).

The second last column, Col. 8, shows the score participants gave to the importance of tracking on a Likert scale of 1 to 7. As one might expect, this is biased towards the high end, with 12 of 21 having scores of 6/7. However, 3 had scores below 5 (P17 - 3, P6 - 4 and P8 - 4) even though they had tracked from 7 to 29 months.

The last column, Col. 9, indicates why people tracked. This is based in analysis of their responses to an open question during the pre-interview stage: "*How and why do you track your physical activity?*". We analysed the free responses and found they fell into two categories which we coded as Goal or Benchmark. The 13 (62%) who gave reasons related to gaining awareness are shown as Benchmark. The other 8, marked Goal, stated they had a goal against which they tracked. Both Benchmark and Goal

[5] https://www.npd.com/wps/portal/npd/us/news/press-releases/2015/the-demographic-divide-fitness-trackers-and-smartwatches-attracting-very-different-segments-of-the-market-according-to-the-npd-group/



1	2	3	4	5	6	7	8	9	10
P#	**Dur**	**SOC**	**Gender**	**Age**	**Occupation**	**Education**	**Tracking Importance**	**Purpose for tracking**	**Fitbit used**
P10	6	M	M	55 - 64	IT Developer	U-Grad	5	Goal	One, Zip, Charge
P6	7	M	F	25 - 34	Part-time	U-Grad	4	Benchmark	Charge HR
P11	8	C	F	35 - 44	IT Support	P-Grad	7	Goal	Flex
P7	9	A	F	25 - 34	Property Admin	U-Grad	5	Benchmark	Charge HR
P18	13	M	M	55 - 64	Retired Military	U-Grad	7	Goal	Surge HR
P14	15	A	F	55 - 64	Academic	Prof	5	Goal	Flex
P21	18	P	M	25 - 34	Student	P-Grad	5	Benchmark	-
P12	18	M	M	25 - 34	Student	P-Grad	7	Benchmark	Flex
P5	21	M	M	45 - 54	Dir of IT	P-Grad	5	Benchmark	Blaze, Surge, Flex
P17	22	A	M	18 - 24	Student	Highschool	3	Benchmark	Zip
P4	23	M	M	35 - 44	Engineer	Prof	6	Benchmark	Surge
P13	26	M	M	55 - 64	Professor	P-Grad	6	Benchmark	Flex, Charge HR
P1	27	C	M	18 - 24	Student	U-Grad	6	Benchmark	One, Zip, Flex, Charge, Charge HR
P3	27	P	F	25 - 34	IT Developer	U-Grad	7	Benchmark	Ultra, Zip
P16	27	A	M	18 - 24	Student	P-Grad	6	Goal	Charge, Surge
P8	29	P	M	35 - 44	Professor	P-Grad	4	Benchmark	One
P19	29	P	M	25 - 34	Researcher	P-Grad	5	Benchmark	Flex
P20	36	M	M	35 - 44	Self Employed	U-Grad	7	Goal	Charge HR
P15	37	P	F	45 - 54	Academic	P-Grad	6	Benchmark	Zip, Charge HR
P9	38	A	F	45 - 54	Manager	P-Grad	6	Goal	One, Charge HR
P2	38	M	M	45 - 54	Manager	U-Grad	6	Goal	One, Charge HR

Table 3.1: Participant profiles, ordered in increasing length of tracking (Col. 2). Col. 1: participant ID. Col. 2: months since first tracker use. Col. 3: the exercise stage of change. Col. 4 to Col. 7: gender, age, occupation and education level. Col. 8: score for importance of tracking from 1 (low) to 7 (high). Col. 9: core reasons for tracking, either to gain awareness (Benchmark) or self-monitor against an activity target (Goal). Col. 10: Fitbit devices worn by participants, including past use. N=21.

purposes are in line with the nature of the feedback available from FitBit, particularly the daily progress and weekly mail summary. But it is notable that so many long-term trackers were Benchmarking.

Overall, we recruited a diverse sample of long-term trackers in terms of duration of use, stage of change, age and gender as well as the classification for the reason for tracking, but a highly educated group that tended to highly value tracking.

3.5.2 How our participants had previously used their activity tracker data

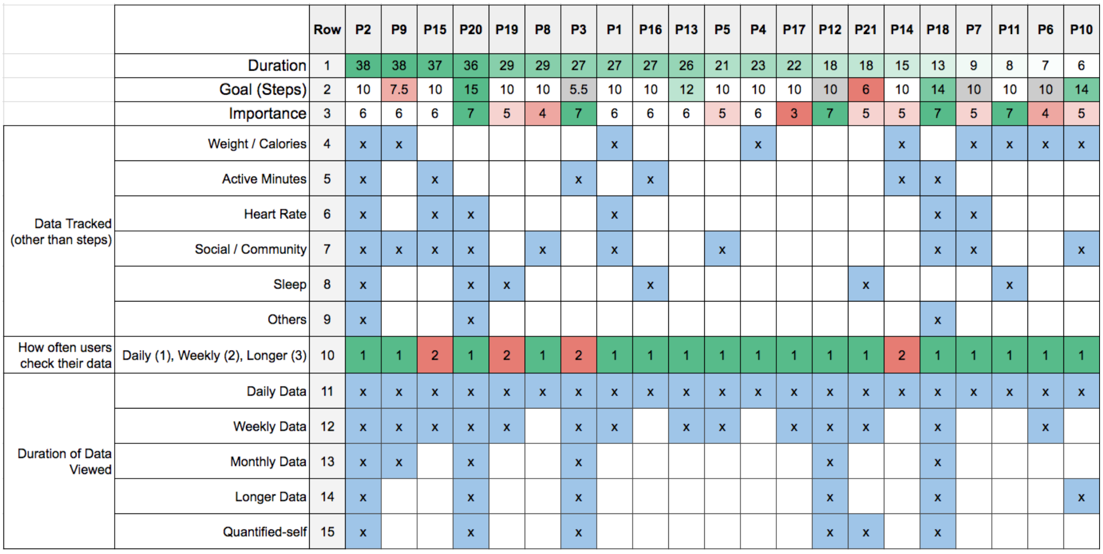
Table 3.2 summarises the pre-interview results on goal targets, importance of tracking and data use behaviours. The table is ordered by tracking duration. We now discuss these.

The default 10,000 steps was dominant.

When we asked an open question about the goals for tracking, 11 participants (52%) stated this was the default 10,000 steps goal. Six had a different step goal and 4 stated that they did not have a step target as their health goal (P3, P7, P12, P6). Out of these, 3 still used 10,000 steps target as a way to benchmark against their actual steps (P7, P12, P6). Three had targets that were based on their knowledge of their actual steps: P3 at 5,500; P9 at 7,500 and P21 at 6,000.

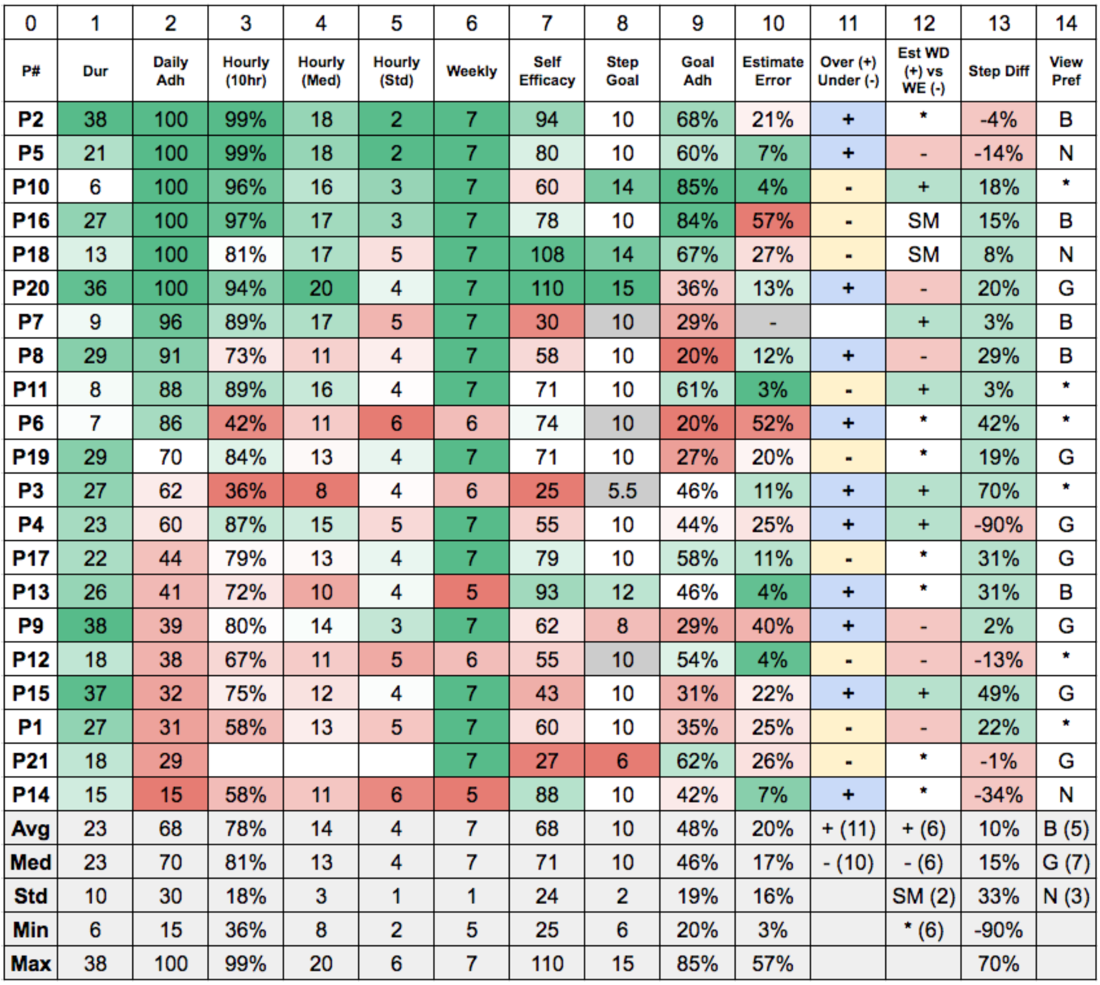
Our participants tracked multiple health data.

In addition to tracking steps, 8 participants (38%) reported tracking weight and / or calories. Losing weight was an important health goal for them and a key reason for tracking health related data. Six participants (29%) reported tracking active minutes, 6 reported tracking heart rate and 6 tracked sleep. Three participants (14%) tracked additional data. P2 tracked cycling and swimming. P20 tracked mood, temperature, food and other exercises. P18 tracked distance, temperature, weekdays versus weekends, time of day of activity.



		Row	P2	P9	P15	P20	P19	P8	P3	P1	P16	P13	P5	P4	P17	P12	P21	P14	P18	P7	P11	P6	P10
	Duration	1	38	38	37	36	29	29	27	27	27	26	21	23	22	18	18	15	13	9	8	7	6
	Goal (Steps)	2	10	7.5	10	15	10	10	5.5	10	10	12	10	10	10	10	6	10	14	10	10	10	14
	Importance	3	6	6	6	7	5	4	7	6	6	6	5	6	3	7	5	5	7	5	7	4	5
Data Tracked (other than steps)	Weight / Calories	4	x	x						x				x				x		x	x	x	x
	Active Minutes	5	x		x				x		x							x	x				
	Heart Rate	6	x		x	x				x									x	x			
	Social / Community	7	x	x	x	x		x		x			x						x	x			x
	Sleep	8	x			x	x				x						x				x		
	Others	9	x			x													x				
How often users check their data	Daily (1), Weekly (2), Longer (3)	10	1	1	2	1	2	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1
Duration of Data Viewed	Daily Data	11	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	Weekly Data	12	x	x	x	x	x		x	x		x	x		x	x	x		x			x	
	Monthly Data	13	x	x		x			x							x			x				
	Longer Data	14	x			x			x							x			x				x
	Quantified-self	15	x			x			x							x	x		x				

Table 3.2: Summary of pre-interview on goal targets, other health related data tracked, data use. Columns are ordered by Row 1, duration of tracking. Top row: participant ID. Row 1: tracking duration in months. Row 2: step goal (grey = participant stated goal not important for them). Row 3: importance of tracking (Likert scale 1-7), Row 4 to Row 9: other data tracked. Row 10: how often checked data, Row 11 to Row 14: duration of the data checked, Row 15: downloaded data for additional analysis. Far right column: %-age of participants with a value for each row. N=21.



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
P#	**Dur**	**Daily Adh**	**Hourly (10hr)**	**Hourly (Med)**	**Hourly (Std)**	**Weekly**	**Self Efficacy**	**Step Goal**	**Goal Adh**	**Estimate Error**	**Over (+) Under (-)**	**Est WD (+) vs WE (-)**	**Step Diff**	**View Pref**
P2	38	100	99%	18	2	7	94	10	68%	21%	+	*	-4%	B
P5	21	100	99%	18	2	7	80	10	60%	7%	+	-	-14%	N
P10	6	100	96%	16	3	7	60	14	85%	4%	-	+	18%	*
P16	27	100	97%	17	3	7	78	10	84%	57%	-	SM	15%	B
P18	13	100	81%	17	5	7	108	14	67%	27%	-	SM	8%	N
P20	36	100	94%	20	4	7	110	15	36%	13%	+	-	20%	G
P7	9	96	89%	17	5	7	30	10	29%	-		+	3%	B
P8	29	91	73%	11	4	7	58	10	20%	12%	+	-	29%	B
P11	8	88	89%	16	4	7	71	10	61%	3%	-	+	3%	*
P6	7	86	42%	11	6	6	74	10	20%	52%	+	*	42%	*
P19	29	70	84%	13	4	7	71	10	27%	20%	-	*	19%	G
P3	27	62	36%	8	4	6	25	5.5	46%	11%	+	+	70%	*
P4	23	60	87%	15	5	7	55	10	44%	25%	+	+	-90%	G
P17	22	44	79%	13	4	7	79	10	58%	11%	-	*	31%	G
P13	26	41	72%	10	4	5	93	12	46%	4%	+	*	31%	B
P9	38	39	80%	14	3	7	62	8	29%	40%	+	-	2%	G
P12	18	38	67%	11	5	6	55	10	54%	4%	-	-	-13%	*
P15	37	32	75%	12	4	7	43	10	31%	22%	+	+	49%	G
P1	27	31	58%	13	5	7	60	10	35%	25%	-	-	22%	*
P21	18	29				7	27	6	62%	26%	-	*	-1%	G
P14	15	15	58%	11	6	5	88	10	42%	7%	+	*	-34%	N
Avg	23	68	78%	14	4	7	68	10	48%	20%	+ (11)	+ (6)	10%	B (5)
Med	23	70	81%	13	4	7	71	10	46%	17%	- (10)	- (6)	15%	G (7)
Std	10	30	18%	3	1	1	24	2	19%	16%		SM (2)	33%	N (3)
Min	6	15	36%	8	2	5	25	6	20%	3%		* (6)	-90%	
Max	38	100	99%	20	6	7	110	15	85%	57%			70%	

Table 3.3: This table summarises all adherence data, self-estimates and their match to actual data, and preferences for goal/ gradient views. Participants (rows) are sorted by daily adherence percentage (Column 2). Column 0: participant ID. Column 1: duration of activity tracker use. Column 2: daily adherence (% of days where users tracked data). Column 3 to Column 5: hourly adherence (% of days with >= 10 hours), median hours per day and its standard deviation. Column 6: weekly adherence (average number of days per week with data). Column 7: self efficacy score (0-110). Cols 8 and 9 show goal aspects: participant's target, actual adherence to that target. Column 10: error of self estimates (i.e., self-estimated daily steps – actual median, as a percentage of actual). (-) means no estimate given. Column 11: shows whether the error in Column 10 was over or under. (blank for no estimate was provided). Column 12: self estimates if more active on weekdays (+), or weekends (-), same (SM) or no estimate given (*). Column 13: actual difference between weekdays and weekends (as % of weekdays). Column 14: preference for goal view (G), gradient (N), both (B) or no preference (N). Summary statistics (mean, median, std, min and max) in the bottom 5 rows. Note: For P21 1-minute data was not available, only daily steps. N=21.

Very few made use of their long-term physical activitydata.

Our participants rated tracking as important (mean 6, on a 7-point scale) shown in (Row 3 of Table 3.2). This is likely attributed to short term self-monitoring, the dominant data available for FitBit. Just 6 participants (29%) reported viewing data that is longer than a month (Row 14). Notably, 1 of the 6 only looked at their yearly step count because they had joined a group which had a 5 million steps per year challenge. Sixteen participants (76%) did not use their activity data for analysis of trends or patterns. While participants had long-term health goals and tracked over the long-term, most have not made use of their long-term physical activitydata.

Only 6 participants (29%) (P2, P20, P12, P10, P18 and P3) reported that they looked at their long-term data. P3 downloaded her data as part of her studies and looked at long-term patterns and trends using a custom application she had built. P12 looked at his monthly data and how he did during summer versus winter, using a spreadsheet. P2 used his data to set step goals. He is the only participant who regularly (monthly) reviews his data in a spreadsheet and actively sets goals based on that analysis. He also monitors his steps between different days of the week to decide when to do more or less. He uses a spreadsheet to do this. P20 checks his steps summaries regularly and sets a goal higher than his average. He also cross references other data (e.g., food, sleep and mood) with his activity data using a third party tool [6] to find correlations. P18 had a history of heart conditions and maintains a spreadsheet of when he walked, how far he walked, what time of day he started, the temperature when he started, average and max heart rate during the time he was walking. He uses excel to do this. He checks his yearly step count for the purposes of tracking against his yearly goal of doing 2016 miles in 2016. P21 had just started to download his activity data at the time of the interview and yet to analyse it.

Use of social engagement.

Almost half of our long-term physical activityparticipants (48%) regularly used social features, such as competing with friends, or participating in online communities. Two participants (P18, P10) were in the 5 million steps challenge and they contribute their activity data to their community. Previous work highlights the benefits of social support and interaction for engaging users over the long-term [Fritz et al., 2014]. This

[6] makesenseofdata.com

paints a profile of the diversity of our participants on this dimension.

Summary.

This sections summarised the participant profiles emerging from the interviews. Most participants tracked a variety of data. Fourteen (67%) used the default goal targets, although 3 considered these unimportant. Most consider tracking as important, but this is primarily for short-term self-monitoring. Over three quarters had not used their accumulated long-term physical activitydata.

3.5.3 Daily and hourly adherence - wearing patterns

We now present the wearing patterns, as summarised in Table 3.3. This table is ordered by daily adherence (Column 2), with the most adherent users first. It shows several wearing adherence measures. Because, self-efficacy about activity may be important, we show it here. Then we present results of goals, participant targets, adherence, accuracy in estimating this and estimates and accuracy about differences in activity over weekends and weekdays. In our analysis, shown in this table, we introduced additional descriptors of adherence, as shown in the column labels. We express daily adherence (days with data) as a percentage of total duration of use (Column 2). Hourly adherence is expressed with 2 calculations: a percentage of days with 10 hours (Column 3) and as median hours of wear (Column 4). The 10 hours percentage was included as it is used in medical science literature as a measure of completeness.

Participants have high daily adherence in weeks with any data, but have diverse patterns of breaks.

Table 3.3 shows the daily adherence %-ages, indicating the proportion of days the person had any data. This ranges from 15% to 100% (mean 68% and std 30%). Low daily adherence indicates many had large breaks. We calculated weekly adherence, the average number of days with data in each week (only using weeks with at least 1 day of data). Our participants had very high average weekly adherence, 7 days per week (std: 1 day) shown in Column 6 of Table 3.3. Of the 21 participants, all had average scores of 7, except 3 (14%) who averaged 6 and 2 (10%) who averaged 5 days per week. This means that in the weeks where they tracked at all, most participants tended to track every day.

To explore breaking behaviour further, we calculated the number and duration of breaks (i.e., no data for 2 days or more). Our participants were diverse on these measures (mean: 18 breaks, std: 25 breaks, max: 105 breaks). The average duration of breaks also varied (mean: 18 days, std: 25 days, max: 98 days). In contrast to the myocardial infarction patient population studied in [Meyer et al., 2016b], our participants tend to break less often but have longer and more irregular break duration (average standard deviation: 45 days).

There were three daily adherence patterns: fully adherent, breaks regularly and breaks in blocks.

We identified three main groups of participant's daily adherence: (a) fully adherent, (b) breaks occur at regular intervals, (c) breaks occur in blocks at irregular intervals. The pattern (a) is clear for the 4 participants with 100% adherence; they have 0 breaks. For others, there was a mix of patterns (b) and (c).

We analysed the data for blocks of continuous data and breaks (Meyer et al [Meyer et al., 2016b] referred to these as breaks and streaks). For example, P8 had a daily adherence of 91% and averaged 7 days per week. He averaged 1.8 breaks per month (3 days per break, std: 2 days) but he had blocks of continuous use that averaged 31 days (std: 27). We characterise his data as being consistent overall with weekly or monthly breaks of a few days. P3 and P13 were similar but with more frequent breaks. They had lower daily adherence overall (62%, 41%) and weekly adherence of 6 and 5 days per week. We can characterise these 2 participant's data as having weekly patterns where they go for few days a week of wear but breaks for 1 to 2 day a week.

By contrast, some participants tended to track in blocks of time with large breaks in between. For example, P1 only recorded data on 31% of days. However, he averaged 7 days per week on weeks when he did record data, had very few breaks (average less than 1 per month) but large and varied break duration (mean: 47, std: 131).

Hourly adherence patterns.

Of our 6 participants with 100% daily adherence, all but P18 had at least 10 hours of data for most days (94−99%) and P18 had 10-hours of wear on 81% of days. These people have high hourly and daily adherence; so they are not prone to the same missing data problems reported in the literature. We studied the break and continuous data patterns for these participants. We found the following: (1) higher hourly adherence (%) was linked to longer periods of continuous and lower number of breaks ($p<0.05$), (2) higher hourly adherence is not linked to duration of breaks. The continuous wear and fewer breaks suggests that participants who are more consistent in wearing their tracker for 10 hours each day are more likely to maintain continuous daily adherence. The second item suggest that when these participants did take breaks, the duration can vary.

Trackers and wearing behaviour.

Col. 10 in Table 3.1 shows the various trackers our participants used over time. We found examples where users reflected on how the tracking device affected their wearing behaviour. For example, P15, who had 32% daily adherence rate, reflected on a large gap in her data in 2013: "*Wow, I didn't wear a FitBit for ages. I was wearing it more in 2013, This would have been Zip FitBit that didn't need charging. So I think that was the difference in wearing behaviour.*". This is particularly interesting because over the long-term, it is conceivable that long-term trackers would transition from device to device where different capabilities and user experiences can result in different wearing patterns. It is then important to have an interface that can make it easy to merge such data.

3.5.4 Goal adherence - activity patterns

We now consider goal adherence and participant's awareness of their activity levels.

Goal adherence patterns.

Column 9 of Table 3.3 shows the goal adherence percentage summary (percent of days a participant achieved their step goals). Averaging across all participants, they achieved their step goals 48% of days. In addition to goal adherence, we also calculated the goal adherence on days where a participant achieved 50% or more of their goal. For example, if a user had a goal of 10,000 steps a day, any days where he reached 5,000 steps or more, this is considered 50% goal adherent. Our participants averaged

50% goal adherence on 80% of days (median: 82%, std: 13%).

We hypothesised a link between higher confidence and goal adherence. We performed a 2 sample t-test between top 10 goal adherence participant's self-efficacy score versus the bottom 11 participants. We found weak significance link between higher self-efficacy and higher goal adherence ($p<0.1$). Interestingly, significance is strong ($p<0.05$) if we removed the 4 participants who stated that they did not have a step goal (P7, P6, P3, P12).

Longer and more adherent use does not appear to improve self-estimates.

During the pre-interview, we asked participants to estimate how many steps they regularly achieve[7] and also whether they are more active on weekends or weekdays. This is because the different routines that many people have on weekends versus weekdays, could impact their activity levels and wearing behaviour. We then compared this with the actual data [8]. Variation in errors in estimates ranged from 3% to 57% shown in Column 10 (mean: 20%, std: 16%). In the table, higher step counts for weekdays are green and this dominates. However, to interpret this we need hourly adherence.

To investigate whether adherence is linked to these estimates, we performed a 2-sample t-test to compare the 10 most daily adherent with the other half, the 11 lower adherence group for estimates of (1) daily adherence (2) hourly adherence and (3) duration. We found no link between daily, hourly adherence and better self estimates. We also found no link when comparing duration with self-estimates.

Column 13 shows large differences between weekdays and weekends and we found no clear relation between adherence and difference in steps between weekdays and weekend. In terms of estimating whether weekdays are more or less active than weekends, out of the 12 participants who provided an estimate, equal numbers gets it right (6 participant) and wrong (6 participants).

In summary, we found no evidence to support the notion that higher daily adherence improves awareness of long-term patterns in activity levels and weekend versus weekdays behaviour.

[7] Grey cell for P7 who declined to provide a step estimate.

[8] For estimating error calculation, we used only days with $>=$ 10 hours to calculate the actual step count as in medical adherence literature.

Goal adherence over or under estimates?

We compared lower versus high goal adherence participants (top 10 in goal adherence % versus bottom 11) We found a tendency for lower goal adherence participants to over estimate their steps compared to higher goal adherence participants with a weak significance (p=0.06). However, we did not find a significant link between goal estimates (Column 9) and error percentage (Column 10 in Table 3.3).

3.5.5 Discovery & insights from wearing patterns

This section draws the qualitative data about the insights that *iStuckWithIt* enabled participants to make. We draw on comments and observations during the think-aloud use as well as the post-interview.

Discovering wearing behaviour.

Participants uniformly commented on the highly visible daily adherence data. Many went on to reflect on this and to explore periods of consistent tracking and breaks, commenting on factors they felt affected their wearing behaviour. For example, P7, had high daily (96%) and hourly (89%) adherence but still commented on periods that were low, "*In Jan / Feb, I forgot to wear it sometimes, or forgot to charge it, I think I'm still getting used to wearing it.*" P4, had lower daily adherence (60%) and commented on the circumstances for low adherence periods: "*I went to join a meeting up in (location) and celebrate the (event), so when I'm in (location) I mostly shut the device off.*" This was especially common for participants with large blocks of breaks.

In the post-interview, P1, when asked whether he learned anything new, answered "*what I learned is the more I wear my tracker, the more steps I have that particular week, I would rather leave my tracker on rather than taking it off right after I get home.*". His response suggests a motivating effect of having more data, reported previously in [Consolvo et al., 2014].

Several participants reported surprise on seeing their hourly adherence levels. P2, who had 99% hourly adherence rate, "*I was surprised to see the number of hours I wore the tracker*". P14, who had 58% hourly adherence but 13 hours of wear (median) commented, "*I am wearing it longer than I think. Interesting, I wouldn't have noticed that*". During the think-aloud, P11 commented: "*18 hours, really, I didn't realise I was* *wearing it for 18 hours. Longer than I thought.*".

Discovering consistency or inconsistency.

Participants also reflected on their wearing behaviour in terms of consistency: P17, who had tracked for 22 months with 44% daily adherence rate, after seeing his data during think aloud period: "*this is interesting, I actually thought I didn't wear my FitBit for a lot longer. I thought I didn't wear it for months*". P19 commented, "*It clearly shows here (early 2014) I've been in the habit of wearing pretty consistently, I've forgotten how consistently I've used it. I really got out of that habit.*".

Making use of missing data.

Where activity tracker data is missing, e.g., due to non wear, participants were still able to reflect on these days in terms of the reason and the context behind non wear as well as what these gaps meant to them. Participants reflected on how much or how little they thought they wore their tracker. P12, with 38% daily adherence rate: "*I've noticed that I don't wear it as much as I thought I did.*". P1, with low daily (31%) and hourly (58%) adherence rates, used gaps as an indicator of activity level. He equated low adherence with lower steps, "*I would say I've decreased in steps recently because I correlate steps directly with wearing of my tracker*".

P8, believed he was more active on weekends (Column 12 in Table 3.3) but he was actually more active on weekdays (29% more - Column 13). During the think-aloud, he reflected on his weekend adherence and implications for accuracy and correct interpretation of the data, "*often weekend days I don't wear it, but if I wear it on weekends, I often have a high number so I'm very active.*".

Some participants reflected on large gaps in their data, to reason about how this reflected their performance. P19 also reflected on a gap in his data and how it affected him, "*looking at 2014, that's when I lost it, at end of July, that's why also there is a gap would make sense in that case. I think the gap really affected me, I got out of habit.*". P6, who had the second lowest hourly adherence rate (42%), commented on her low hourly adherence, "*I just realised, this year, I got a new (pet), I have to wash it a lot and I just take it off. I think it shows less activity because I take it off more.*".

Discovering long-term trends & patterns

When viewing their long-term physical activitydata, our participants often commented that they over or under underestimated their long-term activity levels. For example, P18, who has 67% goal adherence commented, "*I am more active this year, I think because I'm getting stronger and better at it. I started out slow.*". They also commented on the negative patterns or discoveries. P4, who has 44% goal adherence, commented, "*I'm not more active this year, maybe I should do more exercise.*". Some commented on learning something new. P18, who is 100% daily adherent, 81% hourly adherent and regularly reviewed his long-term data still commented on learning something from his data. During the background interview, he reported, "*I don't have any specific hard goals for active minutes, I just make sure I'm up in that range around 120 minutes per day.*". Upon seeing his long-term physical activitydata, "*I seem to be getting well over 100 minutes per day. That's good, I can live with that.*".

Participants also learned about their consistency in their goal adherence over time especially in cases where the data challenged their previous beliefs. P19, (daily adherence 70%, but goal adherence of 27%) commented, "*I really just had not realised how big the shift has been and this really made that clear ... really obvious that I've been very consistent in first half of 2014 and really just dropped off, which I really wasn't aware of*". P8 (daily adherence 91%, hourly adherence 73%) commented, "*I didn't think I was so consistent, consistency between week to week but also across 3 years of data I have.*"

Planning to change goals & strategies.

Ten of our participants (48%) commented on changing goals, making plans or reflecting on strategies for the future. P21, commented, "*I think this data, supports my idea that when I live in a city, I have more opportunity to walk. In the future, this may influence my decision on choosing where to live. I want to be more active so I will probably choose to live in cities.*". P20 after discovering that he was actually less active on weekends compared to weekdays, "*I want to make sure I hit more step goal more on the weekends, walking a little bit more.*". P18 commented, "*I'm very interested in achieving more max days. I've achieved that quite recently, I'm planning to beat that, kind of like competing against yourself.*". P8, had high daily adherence, but with goal adherence of just 20%, commented on changing his goal target, "*I guess*

I should lower it to something more realistic to achieve than the 10K steps per day. Maybe if I change to 8K".

Reflecting on time & context

Most participants reflected on their performance in relation to time (dates, days of week, months and years) as well as context. P8 upon noticing a period of unusually high activity levels during June of 2014, he recalled that he had lent his tracker to his partner. He also reflected: "*In Aug (2014) I was doing (sport event) so I was probably going for runs in the morning, like training for it. When you set the target to 12K steps, you can see those days where I did the runs, and I stopped towards the end of 2014.*".

Participants often reflected on the context and circumstances around specific periods of peaks and troughs. For example, P14 reflected on a period of peak performance, "*Here I was being very fit (early 2015), I was doing dragon boat, swimming bike riding, but clearly it doesn't look as good as here (recent Mar 2016)*". P19, when investigating why he had high lightly active minutes versus low very active minutes on a specific day in Aug 2015, he recounted, "*I actually know what that day was, that day was the day I moved in, in our current apartment. It makes sense it was a long day with a lot of light exercise but few very active minutes.*". P8 examined a peak day, "*25K steps, wow, very high ... I obviously gone for a run or hike on that day. It is sort of interesting looking at it and figuring out what's happening.*". It even triggered a desire to investigate further, "*I would like to have more information for that day. E.g., click on it and see more data. Why it's that value.*".

3.5.6 Feedback on design

We now report participant comments about the *iStuckWithIt* interface and its gradient versus goal views.

What participants "*Liked Most*"

When asked what they "liked most" about the application, 81% of participants commented on clarity and intuitiveness of the calendar visualization. They also liked being able to see the overview as well as details of individual days. A typical comment, from P4 was: "*I like seeing the more comprehensive view, the way the information is* *organised, FitBit gives you the change, more like a vertical bar, ... but your chart gives the overview the whole year, every day, so it is more detailed.*". P15 liked being able to see both overview and details, *"I like the way you can look across month and years even", "I like that you can hover over it certain cells to see details."* Some typical comments about the clarity and aesthetics are: P18: "*It's laid out where you can see the entire week, month and year you're working on. And it goes back all the way to July, so I can look at all of it, complete overview. It's all very straightforward and easy to understand.*". P7: "*I liked the simplicity, in the FitBit versus iStuckWithIt. FitBit is kind of all over the place. I like iStuckWithIt, it's kind of a very simplistic view.*".

Gradient versus goal view.

During the post-interview, we asked users to comment on preferences between goal filtering and gradient views. The summary of preferences is in Column 14 in Table 3.3, with G for Goal, N for gradient and B for both. Of the 15 participants who answered, 47% (7) preferred the goal view, saying it helped them see patterns better. These participants were also able to state activity goals in the pre-interview. Their comments showed that they valued clarity. P20: "*I like goal filtering mode more because the difference is more pronounced. I can really see the blue or the white very quickly and get a quick glance*". P11, "*[gradient view] doesn't give me very useful information as the view before [goal view], because I can't differentiate between different colours ... It's not very easy to see ...*".

Three of the 15 participants (20%) preferred the gradient view. P5 found the amount of white in the goal view was discouraging, "*did not like so many whites. So when goal filtering is off, every box has some colours instead of many white colours.*". The other 2 participants who preferred gradient view said it helped them identify their global peaks better. P14, "*I like the gradient more, I think the gradient tells me more information, the other mode there are just lots of white, I can tell the max better*". P18, "*I prefer without filtering. Because you have your max steps. I'm sure if I exceed the 31k, that will help and spur you to more activity.* "

Five of the 15 (33%) responded that both views were useful, and we observed that they seemed to enjoy switching between them. P7, "*Find it useful to be able to see different views, I like both of them*". P8, "*I like both, a combination is good.*".

Overall, it seems that both views are valuable, with some variation in preferences across participants. To gain insights into how users may use this type of report outside the lab environment. In the post interview, we asked users if, how and how often they would like to view this type of long-term report. Eleven of our participants (52%) preferred to view such data in an email. Five participants (24%) preferred to go to the website regularly or have both. P10 was not interested in the report because he was only interested in tracking towards his daily count and did not see a need for such a detailed report or analysis. P5 preferred to use custom tools he had developed. In terms of frequency, 7 participants (33%) preferred to see this report on a monthly cycle which suggests an appreciation of the long-term nature of this data. Surprisingly, 8 participants (38%) reported a weekly cycle which is very short term. P9 preferred weekly email report but to view the website monthly. Perhaps this is more to do with the existing weekly email summary from Fitbit as 11 (52%) of our participants viewed this regularly. These findings suggest many preferred to receive this type of long-term report as part of regular updates. They prefer longer duration or even on-demand for more interactive engagement in the website.

3.6 Discussion

We now discuss our results in terms of our two research questions, the implications of our work for the design of interfaces onto long term physical activity and future work.

3.6.1 RQ1: How do people currently use their long-term physical activitydata and how does this relate to their actual daily, hourly and goal adherence?

Most of our participants rated tracking as very important. This was to be expected for long term trackers (median 23 months) and their average daily adherence rate was 68%, with about half (10) over 80% daily adherence. Delving deeper into their feedback it becomes clear that this importance is more to do with short term uses such as having a benchmark indicator. We now summarise findings about our participants' use and understanding of their long-term physical activitydata.

People accumulated long-term physical activitydata as a by-product of short term tracking.

This is striking and may well be due to the lack of tools for easily making use of long-term physical activitydata. Although our participants have been using their trackers over a long period of time, most have not been able to harness their long term data. Six (29%) had downloaded their data, then each used different analysis tools for different goals. This is in line with the lived informatics view of personal tracking [Rooksby et al., 2014] where tracking data is often for short term goals.

People have substantial errors in their activity estimates, even when they have high daily and hourly adherence

The average error in step estimates was 20% (median 15%). Somewhat surprisingly, even among the 6 participants with 100% daily adherence (hourly adherence: 81-99%), three had errors of above the overall average (21%, 27% and 57%). While our main goals for *iStuckWithIt* was on overall adherence measures, we also explored differences between weekend versus weekday activity levels. Public health measures of activity are designed to account for this and many people have differences [Tang and Kay, 2016]. Our participants averaged a difference of 10% (median 15%) with wide disparity between them (std: 33%). Of the 12 participants willing to estimate whether they had higher or lower activity levels on weekend versus weekdays, 50% (6) participants got it wrong. Future work could explore ways to provide more details of these differences, both for wearing and activity behaviours. Overall, we found no link between the duration of wear and being more accurate with self estimates. It seems that longer term trackers are no more aware of their long-term physical activitytrends and patterns.

Goal adherence is linked to over- or under-estimates

We examined whether higher goal adherence was linked to higher accuracy in step estimates. Surprisingly, while participants with higher goal adherence are not more accurate in steps estimates, they are more likely to under-estimate their steps and the opposite is true for lower goal adherence participants. It is unclear from our data as to the why this occurs but it may be related to the nature of human memory.

Link between goal adherence and self-efficacy (confidence)

A self-efficacy score was used to measure participants' confidence in meeting activity goals. It is often used in health and education research as an indicator for achieving long term goals. A study of 400 women [Rooney et al., 2003] reported that increased use of activity trackers was linked to self-efficacy. While we did not find a near significant relationship between duration, daily, hourly adherence and self-efficacy, we did find near significant link between higher self-efficacy and higher goal adherence ($p<0.1$). This becomes strong if we exclude the 4 participants who did not have a step goal ($p<0.05$). These results support the link between goal adherence and self-efficacy for people with goals.

Daily adherence patterns

We observed 3 main daily adherence patterns in our participants: (a) fully adherent, (b) regular breaks and (c) large breaks. It was common for our participants to come back to tracking from large breaks, some after months. There was also striking variability in break behaviours, both between individuals and also within individuals. The fully adherent users were quite similar to the 34 myocardial infarct recovery patients in [Meyer et al., 2016b] but our participants were even more consistent in daily adherence. Our work adds to the literature emerging about breaks and consistent daily adherence [Epstein et al., 2016b] and highlights the need to consider them in presenting long-term physical activitydata. Notably, this also points to the poor accuracy of a weekly summary that ignores daily adherence. For a person who has even 1 missing day a week, this mean they underestimate their activity level. Moreover, failure to account for variability in breaks can lead to unpredictable inaccuracies in self estimates.

Hourly adherence patterns

In contrast to daily adherence, hourly adherence for individuals was less varied across individuals (median 81%, std 18%) and within individuals (median 13 hours, std 4 hours). This suggests that our long-term physical activityparticipants tend to be consistent in terms of *how long* they wear their trackers *on days they do wear* their tracker. We also found a positive link between daily adherence and hourly adherence and a link between higher hourly adherence (both percent and actual hours per day) to the length of continuous wearing streaks. This suggests that the higher a user's daily adherence, the higher their hourly adherence over consecutive days. our participants had higher hourly adherence than most of the 237 University students who were given a FitBit for a semester [Tang and Kay, 2016]. This suggests hourly adherence may be a marker of longer term trackers and further investigation is needed.

3.6.2 RQ2: What insights can people gain from our interface which shows daily, hourly and goal adherence over the long term?

Insights about wearing behaviour in daily and hourly adherence

Our participants, on first seeing their data in *iStuckWithIt*, were able to interpret their daily adherence pattern. This also evoked reflective comments on wearing patterns. This is in line with the priority we placed in making daily adherence very visible. This is the first work to report people's reflection on their long term wearing behaviour, an aspect that is important for making sense of long term physical activity. Some found it confirmed their own understanding. But comments from a third of participants indicates insights, either on higher or lower than expected daily and hourly adherence levels.

Inferences about context and factors around notable features in daily and hourly adherence

The choice of calendar layout was intended to give people a broad overview of their data in a format that may enable them to recall relevant context and factors affecting both wearing and activity behaviour. The think-aloud studies confirmed that this was effective. Most participants with periods of low daily adherence reflected on this and recalled reasons and explanatory factors (e.g., holidays, injury).

Insights from goal adherence patterns, steps and active minutes

A key design goal was to make goal adherence visible and the calendar format provided two views. For people who have a goal target and want to see their performance against that, the *goal view* allows them to set two levels of goal adherence. Although our default setting was 100% and 50%, the user's needs may well make other levels meaningful for them. The *gradient view* supports more open exploration. This was intended for people who did not have an explicit goal, and for broader exploration of activity. Both these views enabled people to explore their data in terms of *step goals*, which are entrenched in the FitBit feedback, as well as *active minutes*, which is widely used in public health recommendations (such as [Haskell et al., 2007]). Several participants spoke of their goal adherence in terms of the colour, for example, being pleased to see lots of bright blue, wanting to have more blue or less white.

3.6.3 Lessons for design of interfaces onto long-term physical activitydata

Our study confirms that *iStuckWithIt* met its four design goals. DG1 was to ensure users could readily see *goal adherence* in terms of steps or active minutes per day. DG2 and DG3 were to make wearing behaviour visible, with daily adherence highly visible on the calendar, with hourly adherence available for more detailed exploration of data. DG4, to support reflection about trends, events and context harnessed the calendar layout to help people recall salient factors and context. Our interface's design, with daily adherence seen with goal adherence, enabled all participants to consider these together, reflecting on how well they met their activity goals, but taking account of the data availability and their wearing behaviour. Our interface supports those who want to see granular data, which Epstein et al [Epstein et al., 2016a] found was the case for longer and more consistent users. We offer the following additional insights around our calendar interface for reflecting on long-term physical activitydata.

The calendar supports reflection on factors and context affecting goal adherence

Encoding activity data into a calendar chart enabled users to reflect on their long-term physical activitydata in terms of their context at the time. In addition, the interactive pop-up (or tool-tip) supported detailed exploration. *iStuckWithIt*, enabled participants to reflect on long term trends and discover patterns that mattered to them. Our work confirms the power of the calendar format. This builds on the work of Huang [Huang, 2016] although the integration of activity into a regular calendar is different from our use of the calendar format. There may be synergies in using both, the integrated to raise awareness of activity during the regular use of a calendar and an interface like *iStuckWithIt* to support reflection on long term data.

Wearing behaviour in terms of daily and hourly adherence is valuable.

We introduced these measures of adherence to represent wearing behaviour. This has two key roles. First, it addresses the well documented challenge of interpreting long-term physical activitywith incomplete data. This is what our participants found when they studied hourly adherence to understand apparent changes in goal adherence; they could often discover that they had simply not worn the tracker for a period of time and, with the calendar context, could often reason about the causes. It enabled them to make sense of their data. As physical activity tracking capabilities extend to non-dedicated activity tracking devices such as smart phones and smart watches, we can expect wearing patterns and adherence profiles to change. As such, daily and hourly adherence measures seem even more important as one of the key indicators for comparing activity level data from different devices, time periods and use. In the example of P15, she was more adherent in 2013 because the Fitbit Zip tracker had a much longer battery life compared to more contemporary wrist based trackers that are perhaps more desirable but, at least in her case, more difficult to consistently wear. We demonstrate that by integrating goal, daily and hourly adherence data in a calendar chart design, users can reflect on whether the differences in the visible activity levels are due to changes in behaviour or if they are impacted by incomplete data or perhaps a combination of factors. A second benefit of making daily and hourly adherence visible is that participant could reflect on their wearing behaviour and consider changing it. Indeed, some of our participants reported that they felt motivated to see more coloured cells or wear their tracker longer. This possibility of affecting wearing behaviour is worthy of further investigation. The long term motivational effects of our designs is worth further exploration.

Both goal adherence and gradient views are valuable.

We found that aggregation at the daily activity level in the form of goal adherence is valuable, with the option of gradient and goal views. The former is especially useful for those with striking peaks in activity. These appears to be an effective way to address the challenge of providing an overview of activity data over a long period, both for users who want to think in terms of goal targets and for those who want to do broader exploration. Our goal filtering view design of encoding activity levels in the four colour-codes also proved effective. The facility to alter the thresholds enabled participants to explore their data on multiple values for full and partial goal-adherence. Several participants commented that they preferred the goal filtering view because it was easier to see large peak days which skews the gradient view. Some preferred the gradient view as it helped them see peaks and low periods. Some had negative reactions to seeing white coloured cells (e.g., P5) while it was motivating for others (e.g., P11). Five participants also reported liking the ability to both views. We recommend both goal-adherence and gradient to support different forms of reflection.

3.6.4 Potential roles and uses for *iStuckWithIt*

How might people want to use *iStuckWithIt* within their busy lives? We designed it to support meta-cognitive activities of self-monitoring and self-reflection of long term physical activity. Our lab study clearly demonstrated that participants did engage in these same meta-cognitive activities about their wearing behaviour. Our results point to the meta-cognitive roles of *iStuckWithIt* and how people might want to use it.

Core to all these metacognitive activities are people's physical activity goals. Goal setting theory suggests that effective goals should be "S.M.A.R.T": specific, measurable, attainable, relevant and have a time frame [Locke and Latham, 2002b]. Like our participants, a user could use *iStuckWithIt* to become aware of their long term and more recent activity levels, accounting for changes and patterns over months, seasons and years. This could support setting an attainable goal and they may also plan to return to *iStuckWithIt* to check if they were achieving new goals. For example, for a goal with time frame of 1-month, the user might check their progress at the end of the month. For long term self-monitoring of goals, a subsystem could trigger an email alert when major changes are detected, such as a dramatic decrease in long term activity levels this year compared to last year.

A major life change might also be a trigger to use *iStuckWithIt* to see the effect. For example, P21 concluded that moving from a central location to a suburb correlated with a drop in activity and they stated that they planned to move back. By contrast, P8 discovered that changing location did not have as large an effect on his activity level as he expected. If a user like P21 actually did move house expecting it would affect their activity level, they may want to check the effect using *iStuckWithIt* a few weeks later.

Regular, scheduled use of *iStuckWithIt* could be supported with an email taking them to the interface. This is similar to Fitbit's weekly email service. We asked participants how often they would like to use it, with most opting for weekly email with an option to go to the website for more interactive viewing at a longer frequency. New public health advice is another potential trigger to return to *iStuckWithIt*. For example, many of our participants relied on the default of a 10,000 step goal. If they read a news article claiming it is important to do 10 very active minutes per day, or 60 minutes of moderate activity, they may return to *iStuckWithIt* to discover whether they meet that goal target. This may trigger setting new goals. A scheduled email service, discussed above, might include the option to also receive information about important new public health guidelines. At an individual level, people who have particular medical conditions may benefit from a personalised version of such a service. For example, a recent meta-analysis of exercise for people with cardio-vascular disease points to new understanding of the relative benefits of high intensity and moderate intensity continuous training for this particular population [Liou et al., 2016]. Another role for *iStuckWithIt* is to share it with an advisor, such as a doctor and health coach.

Overall, we envisage that beyond the lab, *iStuckWithIt* use would be linked to people's highly individual physical activity goals. These goals may be explicit, such as increasing to 10,000 steps a day or maintaining current levels. They may be vaguer. Broadly, one class of triggers for using *iStuckWithIt* are scheduled reviews of activity to gain awareness, reflect and self-monitor, then potentially to plan changes. Another important class of potential triggers could come from automated processes, such as reporting new health guidelines or analysis of the data to highlight important changes in activity levels. This might also apply to wearing behaviour if daily or hourly adherence becomes too low for reliable assessment of activity, or with motivational triggers when they continue to use of trackers for extended periods of many days, weeks or months (described as streaks in [Meyer et al., 2016b]) with an acknowledgement reward of seeing more coloured cells in their long term data.

3.6.5 Future Work

Health and well-being are long term endeavours that require personal control and self regulation [Bandura, 2005a]. We could further explore how our interface can support long term SMART goals. We could also explore the effectiveness of different feedback frequencies, triggers and how they can be tailored to users needs and circumstances. A promising direction is how such long term data user interfaces can support interaction between users and their medical practitioners or health advisers. We see opportunities to extend the current interface by integrating with user's email and calendar to access more contextual information (e.g., via a click on calendar cell). Such integration could enable filtering or highlighting capabilities (e.g., show activity levels only on days when I am at work or when it is not raining). We can also provide users the ability to annotate or label days as a way to record or highlight special days and time periods (e.g., training for marathon). In addition, there is an opportunity to design user interface scaffolding for reflection and goal setting which can be useful in learning domains for supporting complex learning goals [Azevedo and Cromley, 2004]. The design of *iStuckWithIt* drew upon physical activity literature to determine that steps and active minutes are meaningful ways to interpret Fitbit data. Adapting it to show other information that sensors now collect, such as heart rate and sleep will also need careful design, based on careful analysis of the relevant literature. Further investigation is needed to examine how and when to show hourly adherence. As mentioned previously, the weekly bar graph may be better used to show other information especially those with consistently high hourly adherence. We can explore options based on triggers or thresholds which may be controlled by users through settings (e.g., highlight or hide incomplete days defined as days with less than 6 hours of data). Finally, we would like to study authentic use of *iStuckWithIt* over the long term, both with healthy people and those with special needs.

3.7 Conclusion

Increasingly, people are amassing long-term physical activitydata. These have the potential to play an important role in reflection, goal setting, monitoring and planning. Currently, there is a gap in understanding the ways that people have used such long term data and an outstanding need for interfaces that enable people to harness that potential.

Our study aimed to examine if designs that taking into account incompleteness can help users extract insights from long-term physical activitydata. We introduced the notion of *adherence* to design *iStuckWithIt*. This uses a custom calendar chart design to show wearing patterns – as *daily and hourly adherence which capture when people wore their tracker and for how many hours* – and embedding activity data, as steps or active minutes, to support reflection on *goal adherence* (on days when they had data, did they achieved their goals). We report a study of 21 long term FitBit users (average: 23 months, 17 > 1 year). It began with an interview about their use and knowledge of their long-term physical activitydata, followed by a think-aloud study with *iStuckWithIt* and a post-interview. The initial interview provides new understanding about long term trackers: their daily adherence showed diverse patterns with some having considerable gaps; but hourly adherence was more consistent; people met their step goals on less than 50% of days with data, but they reached at least half of their daily goal on 80% of days. Surprisingly, we found that the longer term and higher daily and hourly adherence users in our participant population are not more likely to be aware of their step counts and differences between weekday versus weekend. This extends support for the need to investigate interfaces that support reflection on long-term physical activitydata. This work makes two main contributions: new understanding of the ways that people with long-term physical activitydata *have used it and how well they know about it*; the *design and evaluation* of *iStuckWithIt* that demonstrates adherence data inclusion in interface designs can help people gain insights on their long-term physical activitydata even when data is incomplete.

Chapter 4

Designing for Adherence: Scaffolding

Preamble

> Tang, L. M. and Kay, J. (2018a). Scaffolding for an olm for long-term physical activity goals. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 147–156. ACM

This work was published in the International Conference on User Modeling, Adaptation, and Personalization in 2018. It reported on contributions described in section 1.2.3.

Abstract

An important role of open learner models (OLMs) is to support *self-reflection*. We explore how to do this for an OLM based on fine-grained long-term physical activity tracker data that many people are accumulating. We aim to tackle two well-documented challenges that people face, in making effective use of an OLM for reflection. 1. We created a tutorial to scaffold *sense-making* needed to understand the meaning of the OLM. 2. We integrated an interface scaffold to help users *consider key questions* for effective reflection. We report the results of a qualitative think-aloud lab study with 21 participants viewing their own long-term OLM. To evaluate the *tutorial scaffolding*, we split participants into an experimental group, who did a tutorial before exploring the OLM and a control group which explored the interface without the tutorial. To evaluate the *reflection scaffolding*, all participants first explored the interface as they wished. We then provided *goal prompts* to scaffold reflection. Our study revealed that, under lab conditions, the tutorial scaffolding was not needed – all participants in both groups could readily understand the OLM. However, we found that several of the goal prompts were important to help participants consider key questions for effective reflection. Our key contribution is insights into the *design of scaffolding for reflection in a life-long learning context* of gaining insights and setting goals for physical activity.

4.1 Introduction

There is a growing body of work that aims to create Open Learner Models (OLMs) to support meta-cognition [Guerra, 2016, Bull et al., 2016, Long and Aleven, 2013, Desmarais and Baker, 2012, Tabuenca et al., 2015]. Much OLM research has been in the context of formal education. However, this important role of user models is also important in *life-long* and *life-wide* learning. In particular, one key role of an OLM is to support *self-reflection* [Bandura, 2005b, Deci and Ryan, 2000, Bull and Kay, 2010, Desmarais and Baker, 2012, Bull and Kay, 2016]. This is especially important for achieving very long-term goals, such as achieving and maintaining healthy levels of physical activity [Barua et al., 2014, Marcengo et al., 2016, Kay, 2016]. OLMs can play several important roles, including support users' curiosity about their data, allowing for playfulness in tracking style [Rapp and Cena, 2016], learner trust [Ahmad and Bull, 2009] and several broader important meta-cognitive activities [Bull and Kay, 2013, Feyzi-Behnagh et al., 2014, Duffy and Azevedo, 2015, Bull and Kay, 2016].

Our work aims to take insights from OLM research into the area of *personal informatics*, where emerging sensor technologies enable people to collect considerable personal data. We focus on the goal to harness data from worn sensors about physical activity. Such sensors are becoming ubiquitous with the emergence of dedicated devices such as the Fitbit [1] as well as through ambient tracking via non-dedicated devices such as smartphones [2] [3]. Our work takes an OLM perspective, first to transform long-term physical activity data into a user model, then to create an OLM interface, called *iStuckWithIt*, to support self-reflection on the user model.

In this paper, we consider two research questions about the scaffolding for self-reflection using *iStuckWithIt*:

- Do users need a scaffolding introductory tutorial to self-reflect using *iStuckWithIt*?
- Do users benefit from a reflection scaffold to systematically self-reflect on core long-term goals represented in the OLM?

To explore this, we conducted a lab study where 21 existing long-term physical activity trackers were asked to use *iStuckWithIt* [Tang and Kay, 2017], with 2 additional scaffolding elements: *tutorial* introduction and *goal prompts* for reflection. The tutorial scaffolding asks users to review the data of 2 hypothetical users, with data that highlights critical features of the dashboard. The goal prompt scaffolding is a side panel (pop-up) that asks users to answer 5 questions about their goals and their behaviour including whether they are achieving their goal, whether they should change their goal and to consider differences between when they are at work and not at work, weekend and weekdays or on holidays. These questions prompt users to considering their goal setting as well as how environmental and temporal factors that are known to affect physical activity, as documented in health literature [Cadmus-Bertram et al., 2015a].

The next section reviews related work followed by the study design and results. We conclude with a discussion of the findings and lessons learned for future designs.

[1] fitbit.com

[2] https://www.apple.com/au/ios/health/

[3] https://support.apple.com/en-au/HT203037

4.2 Background

This section positions our work in relation to three bodies of previous research. First, we build on OLM research, where a user model is made available to the user to support goals such as self-reflection. Then we introduce the largely independent work on personal informatics, including an overview of the design of the basic *iStuckWithIt* interface. The third key strand is on meta-cognitive scaffolding for OLMs. We then introduce the main *iStuckWithIt* interface and explain the goals of our work in terms of the new contributions we aim to achieve.

Open Learner Modelling has a long history, beginning with the recognition that a user model (also called student or learner model) could be made available to the user [Kay, 1994, Bull et al., 1995]. An OLM could serve several roles, including the learner interacting to negotiate or argue about the user model [Bull et al., 1995], supporting user control over their personal data [Kay, 1994] and for metacognitive processes of self-reflection, self-monitoring and planning [Bull and Kay, 2010, Desmarais and Baker, 2012, Bull and Kay, 2016]. There has been considerable research on the ways to present learner models, comparing various forms [Bull and Kay, 2016, Guerra-Hollstein et al., 2017]. There is also a body of studies that have demonstrated their effectiveness for learning in formal educational contexts [Mitrovic and Martin, 2002, Mitrovic and Martin, 2007, Desmarais and Baker, 2012, Long and Aleven, 2013].

While OLM research has been largely concerned with formal educational settings, emerging sensor and mobile technologies have led to Personal Informatics research [Li et al., 2010] and the similar Quantified Self movement in the broader community[4]. These communities also aim to create useful representations of users, available in interfaces that have similar goals to OLMs. This community has demonstrated that, while people see the potential value of such data for self-reflection, current tools fail to support this well [Li et al., 2011, Rapp and Cena, 2016, Rooney et al., 2003, Tang and Kay, 2017, Choe et al., 2017, Gouveia et al., 2015]. Indeed, there is a growing body of evidence that points to a lack of perceived usefulness of long-term tracker data [Rapp and Cena, 2016, Fritz et al., 2014, Rooksby et al., 2014]. In personal informatics, the user models need to be designed to represent aspects of user's goals, linking the available sensor data to those goals. A key problem

[4] http://quantifiedself.com/

in creating the model, and associated OLM interfaces, relates to problems in the accuracy of the data due to incompleteness. For example, a worn activity tracker only gives reliable data when the user wears it and this should be considered in reasoning about the user's activity level and modelling their goal achievement. Incomplete data compromises the usefulness of tracking. People can lose confidence when they are confronted with gaps or incorrect reports due to gaps [Rapp and Cena, 2016, Bentley et al., 2013]. Failure to account for, or recognise, incomplete data can mean that people consider the data not to be useful which has been reported in recent years [Rapp and Cena, 2016, Lazar et al., 2015, Shih et al., 2015, Fritz et al., 2014, Yang et al., 2015, Harrison et al., 2015, Elsden et al., 2015]. A similar problem has been identified for OLMs for formal learning, with the need to represent the uncertainty in the model [Epp and Bull, 2015, Al-Shanfari et al., 2016].

While the ideal OLM interface would be readily understood by the user, in practice this may be difficult to achieve. Even with a quite simple skillometer, consisting of just seven bars [Long and Aleven, 2011], there were challenges in both understanding the model and the meaning of the components display as well as in reflection. Self-evaluation is especially important for achieving the very long-term goals, such as achieving and maintaining healthy levels of physical activity [Barua et al., 2014, Marcengo et al., 2016, Kay, 2016]. OLMs can play several important roles, including supporting users' curiosity about their data, allowing for playfulness in tracking style [Rapp and Cena, 2016], facilitating learner trust [Ahmad and Bull, 2009] and several other important meta-cognitive activities [Bull and Kay, 2013, Feyzi-Behnagh et al., 2014, Duffy and Azevedo, 2015, Bull and Kay, 2016].

This work explores the role of scaffolding for the *iStuckWithIt* interface. The design of this interface and the nature of insights people made when using it have been previously reported [Tang and Kay, 2017]. We now briefly introduce that version of the interface, as shown at the left of Figure 3.1. Broadly, the design is based on a calendar visualisation. The labels A-H illustrate key features. A marks the drop-down menu to select the class of goal the user wants to see; those in the study are steps per day, count of active minutes per day and distance walked per day. The main interface is marked B for a period when the user has data about their activity levels (with all cells either white or shades of blue) and C when there was a period with no data because the user did not wear their tracker in that period (grey cells). The figure shows the interface configured for a goal target of 10,000 steps a day and only cells exceeding this are bright blue. The configuration in the figure sets a 50%, or 5,000 step threshold for the lighter blue and then white indicates days that have data (>= 1 step) but less than 5,000 steps. The bars marked with D were designed to help the user take account of the impact of their actual wearing behaviour on the results shows. The bars show the average number of hours per day the user wore the tracker in each week. When this is low, as in the case of weeks nearest the D, the results are based on just the limited data that is available. In the figure, the user has clicked on the cell near E to see more details for that day. The upper middle is the configuration section, labeled F. This enables the user to change the thresholds for the goals. The right part of the figure, G, is the reflection scaffolding that is the focus of this paper. H enables the user to alter the display from goal oriented, as in this figure, to a gradient.

In summary, there has been considerable research in OLMs, especially in formal educational settings. There is a growing body of work in personal informatics and broader community interest in Quantified Self. Both have identified a key challenge for effective use of personal data – although they have not used the term, OLM, they highlight the need for scaffolding to help people make sense of complex collections of personal data, user models, so as to support their self-reflection. Our work tackles this problem. This paper goes beyond our previous report of *iStuckWithIt*'s design [Tang and Kay, 2017] as we now describe the study of the two forms of scaffolding we explored for the *iStuckWithIt* OLM interface: the tutorial scaffolding to introduce the interface and the reflection scaffolding.

4.3 Study Design

Our two research questions were:

- RQ1: Tutorial scaffold: Do users need a scaffolding introductory tutorial to self-reflect using *iStuckWithIt*?
- RQ2: Goal prompts scaffold: Do users benefit from a reflection scaffold to systematically self-reflect on core long-term goals represented in the OLM?

This section first describes the overall design of the study and then the detailed design for each research question.

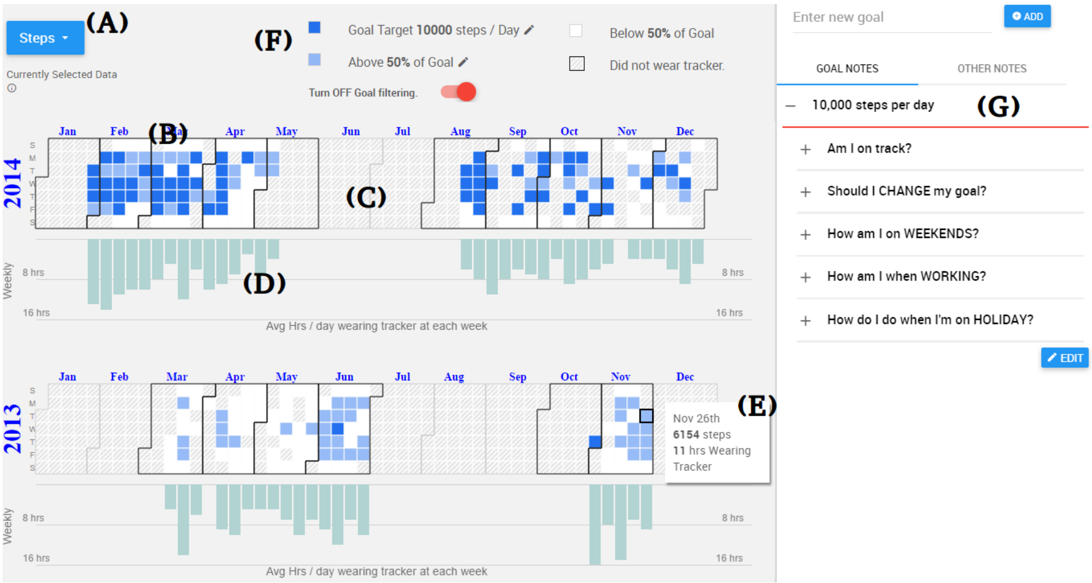


Figure 4.1: OLM to support reflection on achievement of long-term physical activity goals. (A): Drop down menu to select between datasets (i.e., steps, active minutes, distance). (B): Calendar visualisation, with colour showing activity level on each day – dark blue means 10,000+ steps, light blue >5,000 but <10,000 steps, white has >0 but <5,000 steps. (C): Grey striped cells are days with no data. (D): Bar graph showing average wear-time in hours per day for each week. (E): Pop-up showing additional details of a particular day / cell. (F): Configuration – to adjust the thresholds for colouring of the cells. (G): Goal prompt scaffolding questions.

We recruited 21 long-term Fitbit users, people who had collected at least 6 months of personal physical activity data. We then conducted a between-subjects lab study in terms of the first research question. This study session had the following stages

1. Nine participants (9) worked through the scaffolding tutorial, described below.
2. All participants were asked to explore the main interface in *iStuckWithIt*. We asked them to *think-aloud* [Nielsen, 1994], explaining what they saw, understood and their insights.
3. We then asked all participants to consider the reflective questions, labeled (G) in Figure 4.1.
4. Finally, we interviewed them on their experiences of viewing their own data and what insights they learned.

In the next two sub-sections, we present the motivation and design of the tutorial

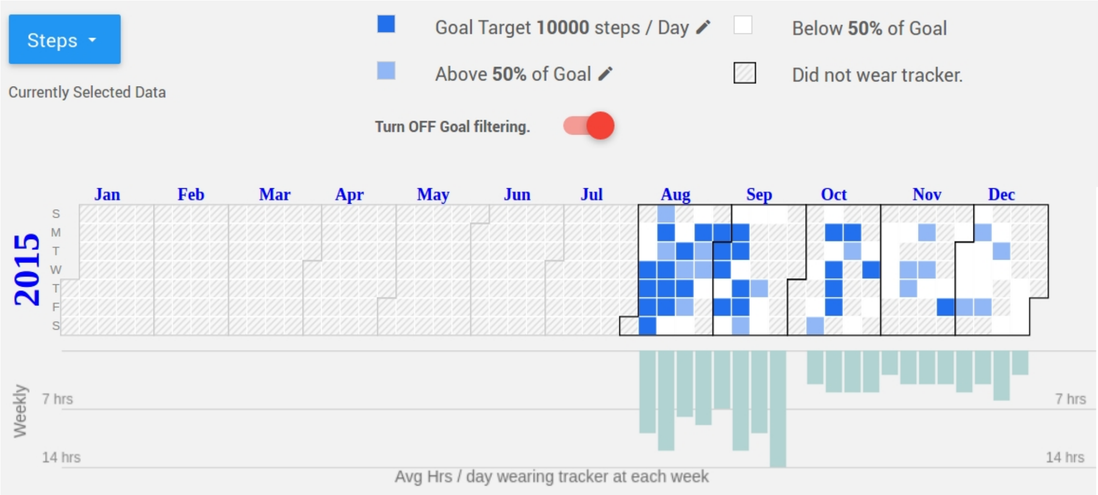


Figure 4.2: Data for hypothetical user Alice where she started using data in August until Dec. Her daily wear-time declined after September especially on weekends and wear-time is also lower but consistent.

and goal prompt scaffolding. We also explain the study design to evaluate each of them.

4.3.1 RQ1: Tutorial Scaffolding

While some long-term physical activity trackers do use their tracker over the long-term, most fail to make use of their own long-term data [Tang and Kay, 2017, Fritz et al., 2014, Rapp and Cena, 2016]. This means that our study design should account for the likelihood that the *iStuckWithIt* interface would provide the first opportunity for participants to see their own *long-term* goal performance for physical activity, in terms of steps, active minutes and distance. We anticipated that participants would benefit from a tutorial that introduced them to *iStuckWithIt*.

To evaluate this, we prepared a tutorial, based on a set of exercises to explore the *iStuckWithIt* OLM for two hypothetical users, Alice and Alex. These provided carefully designed datasets which highlighted key aspects that the interface was intended to enable people to understand about their own long-term physical activity model. We asked 9 of our participants to complete this prior to seeing *iStuckWithIt* with their own data. The steps in this tutorial were:

1. Participants were told that Alex and Alice's each had a goal of "at least 30 very active minutes per day".
2. Participants were asked to consider whether Alice achieved her goal or not.
3. They then did this for Alex.
4. In each case, the experimenter allowed the participant to explore the OLM, thinking aloud to explain how they interpreted it.
5. At the end, if participants had failed to see and understand key features, the experimenter explained them.

Figure 4.2 shows the data for Alice who started using her Fitbit in August 2015. Key features are:

1. The first month of tracking had quite high tracker use - few grey cells;
2. In mid-September, there is a 2-week gap in tracking - all grey cells;
3. After this, there are many grey cells, reflecting days Alice did not wear her tracker, especially on weekends.
4. Consistently higher wear time in August and September and then consistently lower hourly wear-time after September.
5. Scaffold users to reflect or speculate on the *potential causes* for this change from Aug/Sept to afterwards.

Figure 4.1 shows Alex's data which highlights the following:

1. Low physical activity levels during weekend compared with weekdays.
2. Large gaps of several months (blocks of grey cells) between wearing activity tracker.
3. Overall inconsistent hourly wear-time and some periods with low wear-time.

We recorded observations and participant comments in their think-aloud exploration of the interface. If, after some exploration, the participant did not notice key aspects, they were prompted about them. We also recorded whether the user commented on how missing data could have affected accuracy of the step counts as well as comments around wearing behaviour of Alice and Alex.

The aim of the tutorial scaffolding study is two-fold. First, we wanted to discover which features of the long-term physical activity tracker data are easily understood and which are not. We also wanted to see the impact of the tutorial, to learn whether participants who had the extra learning scaffolding were better able to make sense of their own data than those who did not do the tutorial.

4.3.2 RQ2: Goal prompts scaffolding

After participants had finished exploring the main *iStuckWithIt* interface with their own data, we asked them to open the goal prompt scaffolding, labelled G at the right of Figure 4.1. This was designed around two core forms of reflection:

1. Reflect on *goal achievement* and consider reviewing the *goal setting* – the first two questions, marked +.
2. Reflect on *factors affecting goal achievement* – the last 3 prompts, also marked with +, about weekend (versus weekdays), holidays and work versus non-work.

The benefits of scaffolding or support for self-regulation skills such as self-monitoring, goals and goal setting are well documented [Bandura, 2005a, Azevedo and Cromley, 2004]. Also, previous work using goal setting as a strategy to promote health and physical activity behaviour change has demonstrated the potential of this approach [Strecher et al., 1995, Shilts et al., 2004]. The first category of our questions aims to remind users to reflect on their goals and goal setting. The second category of questions called for the participant to consider factors that are known to be important for people's physical activity levels. Health studies have consistently indicated differences between activity levels on weekends versus weekdays [Behrens and Dinger, 2005, Fairclough et al., 2014, Tang and Kay, 2016]. Moreover, previous studies have shown that by helping users consider such questions can support reflection on activity tracker data [Epstein et al., 2014]. Therefore, our scaffolding design was based on literature indicating the benefits of reminding users to consider the context of their activity levels and behaviour is likely to be useful.

In addition to our prompt questions, we did consider others. For example, studies of existing users of physical activity trackers have reported that such users are quite interested in peaks and lows [Fritz et al., 2014, Li et al., 2012]. On the matter of influencing people to change tracking behaviour, Epstein et al. [Epstein et al., 2016b] found potential value in using visualisations for encouraging users to return after a long gap in tracker use. While these are potentially useful, we did not include them as our focus was of studying whether people could make sense of their data at the OLM interface and reflect on important features associated with learning about their own long-term goal achievement.

4.4 Results

In this section, we first introduce the participants of our study then we discuss the results around the two research questions.

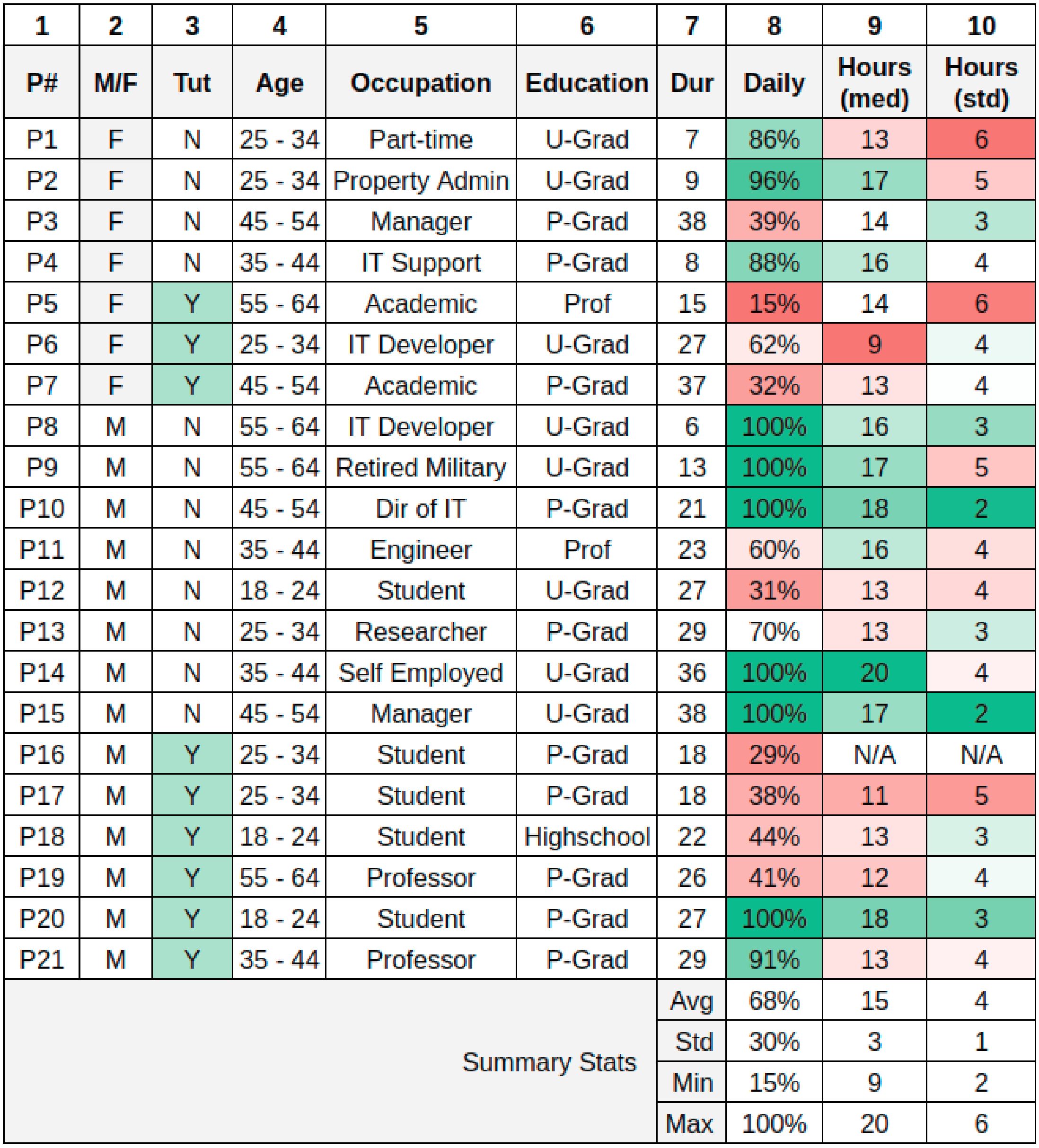
4.4.1 Participants

Table 4.1 presents details of our participants, ordered by gender in Col. 2, then by scaffolding condition Col. 3. This shows 9 participants did tutorial (Y) and 12 did not (N). In terms of background, many participants were highly educated and several worked or studied in IT, shown in Col. 5 and Col. 6. This group may have higher literacy and skills levels in data analysis than a more general population. This is similar to participants in other qualitative studies of existing long-term personal trackers [Fritz et al., 2014]. There were more male participants (14) than female (7). Our participants ages are spread across the age groups shown in Col. 4 with 25-34 being the largest group at 6, the lowest 18-24 (3) and 4 participants in the others. Our demographic is similar, in terms of age and gender, to the population of personal activity tracker users and wearable technology adopters [Endeavour, 2014]. The duration of tracker use varied from 6 to 38 months shown in Col. 7 (average: 23 months).

Col. 8 shows the %-age of days with at least 1 step. This varied widely (min: 15%, max: 100%, average: 68%, std: 30%). Col. 9 is the wear-time (number of hours per day users with $>=$ 1 step recorded). Our participants generally had high wear-time (min: 9 hours, max: 20, average: 15, std: 3) and Col. 10 shows the standard deviation (min: 2, max: 6, average: 4, std: 1). Col. 9 and Col. 10 show that while overall wear-time is high there is large variation both between and within individuals.

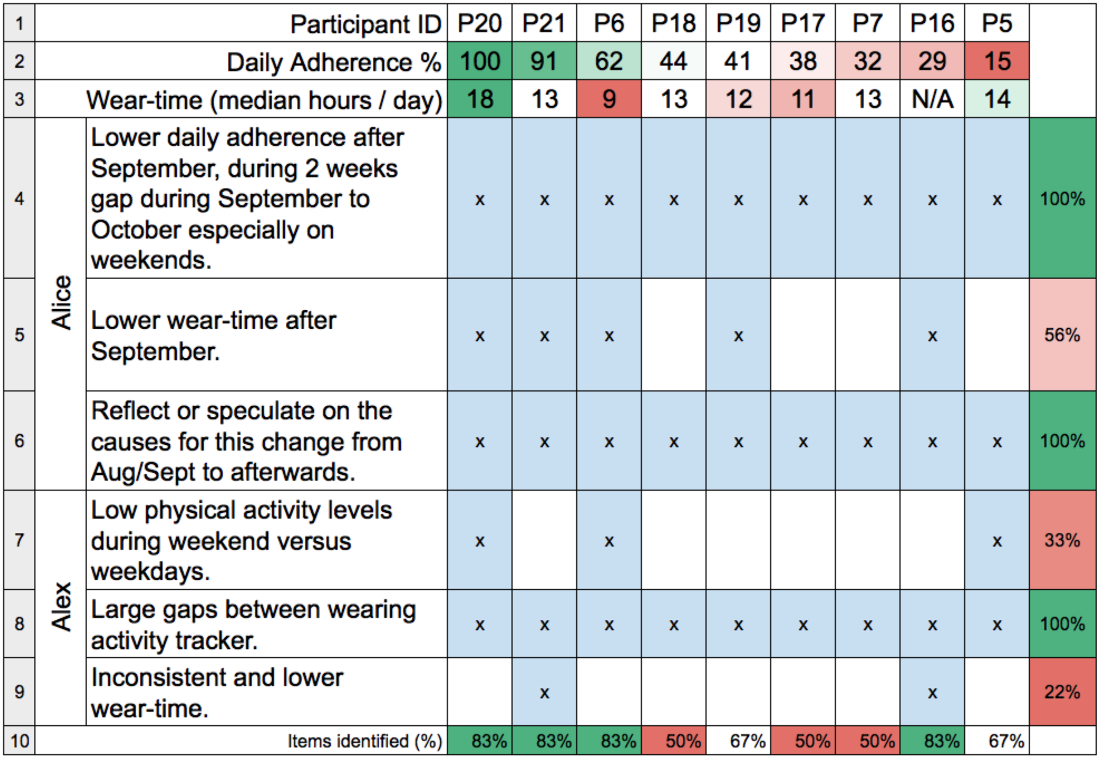
Overall, our participants had very wide differences in consistency of days with tracking – the %-age of days with any data. For example, 6 participants (P8, P9, P10, P14, P15 and P20) had 100% of days with data - meaning they wore their tracker every day in the period from the first day to the last in the dataset. These participants also averaged higher wear-time within each day, recording between 16 and 20 hours per day.

Table 4.1: Participant profiles grouped by gender. Col. 1 participant identifier. Col. 2 gender. Col. 3 whether participant is in the tutorial scaffolding condition. Col. 4 to Col. 6 participant age, occupation and education. Col. 7 duration in months of tracking data (first to last day with data). Col. 8 %-age of days with at least 1 step Col. 7). Col. 9 and Col. 10 the median and standard deviation of wear-time (number of hours per day with at least 1 step). The last 4 rows are summary statistics over all participants (average, standard deviation, min and max) of Col. 8, Col. 9 and Col. 10. N=21



1	2	3	4	5	6	7	8	9	10
P#	**M/F**	**Tut**	**Age**	**Occupation**	**Education**	**Dur**	**Daily**	**Hours (med)**	**Hours (std)**
P1	F	N	25 - 34	Part-time	U-Grad	7	86%	13	6
P2	F	N	25 - 34	Property Admin	U-Grad	9	96%	17	5
P3	F	N	45 - 54	Manager	P-Grad	38	39%	14	3
P4	F	N	35 - 44	IT Support	P-Grad	8	88%	16	4
P5	F	Y	55 - 64	Academic	Prof	15	15%	14	6
P6	F	Y	25 - 34	IT Developer	U-Grad	27	62%	9	4
P7	F	Y	45 - 54	Academic	P-Grad	37	32%	13	4
P8	M	N	55 - 64	IT Developer	U-Grad	6	100%	16	3
P9	M	N	55 - 64	Retired Military	U-Grad	13	100%	17	5
P10	M	N	45 - 54	Dir of IT	P-Grad	21	100%	18	2
P11	M	N	35 - 44	Engineer	Prof	23	60%	16	4
P12	M	N	18 - 24	Student	U-Grad	27	31%	13	4
P13	M	N	25 - 34	Researcher	P-Grad	29	70%	13	3
P14	M	N	35 - 44	Self Employed	U-Grad	36	100%	20	4
P15	M	N	45 - 54	Manager	U-Grad	38	100%	17	2
P16	M	Y	25 - 34	Student	P-Grad	18	29%	N/A	N/A
P17	M	Y	25 - 34	Student	P-Grad	18	38%	11	5
P18	M	Y	18 - 24	Student	Highschool	22	44%	13	3
P19	M	Y	55 - 64	Professor	P-Grad	26	41%	12	4
P20	M	Y	18 - 24	Student	P-Grad	27	100%	18	3
P21	M	Y	35 - 44	Professor	P-Grad	29	91%	13	4
Summary Stats						Avg	68%	15	4
						Std	30%	3	1
						Min	15%	9	2
						Max	100%	20	6

Table 4.2: Table showing whether users identified each of the notable items as part of their tutorial scaffolding. Row. 1 participant ID. Row. 2 %-age of days with at least 1 step. Row. 3 median hours per day with >=1 step. Row. 4 to Row. 6 3 notable items in the Alice tutorial. Row. 7 to Row. 9 notable items in the Alex tutorial Row. 10 notable items identified by each participant as %-age of all 6 such items. N=9.



1		Participant ID	P20	P21	P6	P18	P19	P17	P7	P16	P5	
2		Daily Adherence %	100	91	62	44	41	38	32	29	15	
3		Wear-time (median hours / day)	18	13	9	13	12	11	13	N/A	14	
4	Alice	Lower daily adherence after September, during 2 weeks gap during September to October especially on weekends.	x	x	x	x	x	x	x	x	x	100%
5		Lower wear-time after September.	x	x	x		x			x		56%
6		Reflect or speculate on the causes for this change from Aug/Sept to afterwards.	x	x	x	x	x	x	x	x	x	100%
7	Alex	Low physical activity levels during weekend versus weekdays.	x		x						x	33%
8		Large gaps between wearing activity tracker.	x	x	x	x	x	x	x	x	x	100%
9		Inconsistent and lower wear-time.		x						x		22%
10		Items identified (%)	83%	83%	83%	50%	67%	50%	50%	83%	67%	

In contrast, the 6 participants with lowest consistency in wearing the tracker had between 15% and 39% of days that had any data (P6, P10, P2, P17, P20 and P18) and their median wear-time was 9 to 16 hours per day. Since in medical research 10 hours of wear-time is considered sufficient for meaningful data [Migueles et al., 2017], even those who averaged 9 hours (P6) may have acceptable data quality. Interestingly, these 2 groups of users had similar wear-time variation indicated by the standard deviation (i.e., variation in the number of hours per day with at least 1 step) between averaging 4 hours for lower daily adherence users and 3 for highly daily adherent users.

These participant statistics indicate that while there are large variations in the *number of days with data* (%-age of days with at least 1 step) all participants had high wear-time.

4.4.2 RQ1: Tutorial Scaffolding

Table 4.2 shows which notable item each participant identified for hypothetical users, Alice from Row. 4 to Row. 6 and Alex, from Row. 7 to Row. 9 during the tutorial scaffolding condition. All participants identified at least 50% of these but none could identify all items.

Our participants readily identified whole day wearing patterns, distinguishing days with any data (appearing as white or light or dark blue cells) against days with no data (grey cells). In the Alice condition, all participants observed that she had more missing days after September (Row. 4) and all commented or speculated on the causes for this (Row. 6). For example, P20 commented on Alice's gap in data during September, *"Stopped wearing in September, October returned, maybe had gone on holiday"*. In the Alex condition, all participants commented on the large gaps between periods of more consistent tracking over the 2 year period (Row. 8). They also commented on Alex being more consistent in wearing his tracker on more days in the second year.

The notion of the wear-time in terms of *number of hours with any data in a day*, was harder for our participants to discover. Row. 5 shows that 4 participants (44%) did not comment on the low wear-time for Alice after September. Moreover, only 2 participants (22%) noticed the drop in Alice's wear-time (Row. 5) as well as the inconsistent wear-time across different months and years in Alex's data (Row. 9). P16 commented when viewing Alice's data, *"I see she is not always using her tracker, especially in the last few months. Only 4 - 5 hours per day"* and when viewing Alex's data *"He used to wear the tracker more longer than recently"*. P21, when viewing Alice's data, *"Towards the end, not only was [Alice] less active but also wearing it for shorter periods"*. When viewing Alex's data, he commented, *"Alex is also more consistently wearing the tracker in the second year."*.

Interestingly, most participants (3 of 4) who commented or reflected on wear-time also commented on their own wear-time when viewing their own data during the think-aloud. (The outlier is P16, the sole participant who did not have wear-time data).

Notably, several participants in the *control condition* – who did *not* do the tutorial scaffolding, also commented on wear-time in their own data. For example, P12 commented on how the hours of wear (wear-time) may affect his activity levels, *"If I wear my tracker longer, it will track more of my activity. I take off my tracker when I get home. So maybe I should just leave my tracker on myself at home as well to get more steps, little steps I do walking around home"*. P4 was surprised at the number of hours she wore her tracker, *"I was surprised I was wearing it for 19 hours"*. When investigating a day that had very low steps as well as hours of wear, P5 commented, *"I thought I wore it for longer that day, it could be that I bumped it and turned it off. I think it happens sometimes when I bump it"*.

Row. 7 indicates that only 3 participants (33%) noticed the substantially lower steps and lower days with any data on weekends. One of the three who did, P7, reflected on the differences between Alice and Alex, commenting, *"Alex didn't wear it on weekends, which I didn't notice on Alice"*.

Row. 2 and Row. 3 of Table 4.2 shows each participant's own daily %-age of days with data (Daily Adherence) and wear-time against the number of notable items in each tutorial scenario. This shows the completeness of their own data, so we can see this against the notable items they identified. For example, P6 had the lowest median wear-time (9 hours per day) and P5 had the lowest daily adherence (with just 15% of days with data). During the scaffolding tutorial, we found several cases where users related their own experiences and thinking when viewing these 2 hypothetical user's data. For example, P5 explained their low, 15% of days with data, was due to significant medical conditions, and commented *"Sometimes she doesn't wear her tracker, so she's like me"*. P17 who had 38% daily adherence commented, *"I can see Alex is achieving his goal. But as time passes, he started to drop in his wear. In the beginning, I think he is more motivated, like myself as well."*. When looking at Alice's data, P17 commented, *"I think people forget to wear or perhaps it doesn't match the *fashion”*. Interestingly, when he explored his own long-term data, after the tutorial, he explained that he stopped wearing his Fitbit tracker because his did a sport that required a wrist based apparel that prevented him from wearing his Fitbit. His earlier speculation on the reason for Alice’s missing days could be due to fashion concerns, is similar to this reasoning. P21 commented, *“I can see there are some days where Alex forgot to wear it or forgot to track properly – maybe she was on holiday or something”*. When analysing his own long-term tracker data, P21, who has relatively high daily adherence (91%) commented on the days he did not wear the tracker due to holidays and work travels. P18 who has daily adherence of 44% also commented on the motivational effect of adherence and missing data, *“She was very consistent but she stopped. I think it’s discouraging to have so much missing data.”*.

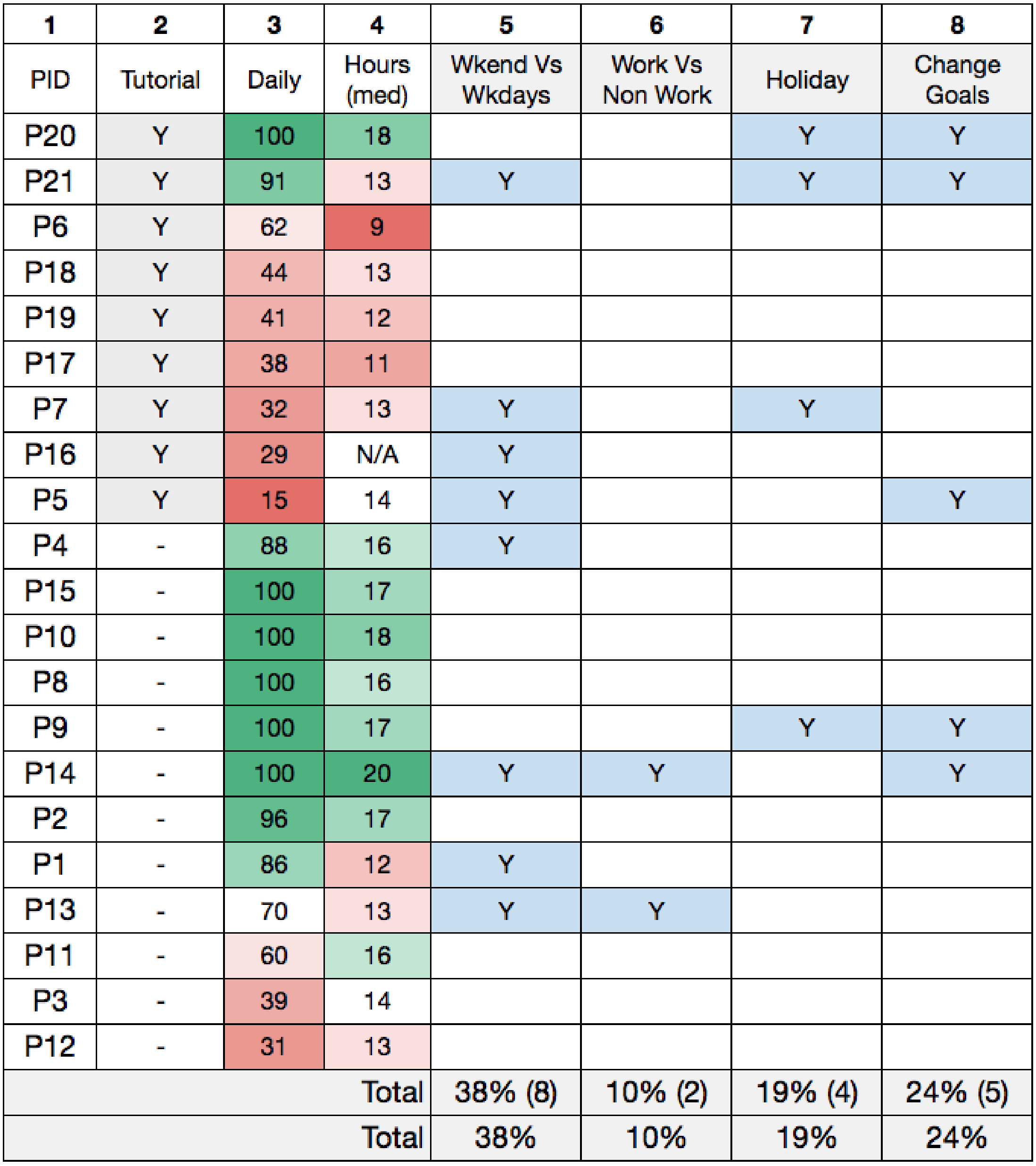
This section summarised results for the first research question which explored whether the tutorial scaffold was needed and useful. Comparing the participants’ understanding of their own *iStuckWithIt* OLM, we observed that participants in both the tutorial and control conditions could understand the main features in terms of *days the goals* were met and how to interpret the display of days with any data. Both groups also performed similarly on wear-time (median hours of wear per day with >1 step), with most people tending to miss this aspect and similar levels of awareness of it between the conditions. Overall, both conditions performed similarly.

4.4.3 RQ2: Goal Prompts

We now consider the results for the scaffolding for reflection. While the tutorial was done only be the 9 participants in the tutorial group, all participants then used the interface to explore their own data. When they had finishing doing this, the experimenter asked them to open the scaffolding section of the interface to consider the questions intended to help them consider and reflect on key features. Table 4.3 summarises the new insights that participants gained in this scaffolded reflection phase. Col. 5 to Col. 7of the table show where the goal prompt enabled a participant to gain new insights ***in addition*** to those the experimenter recorded them as making already.

When asked about weekend versus weekdays, 8 participants (38%) identified new insights – shown in Col. 5. For example, P4 commented, *“I’m not doing very well on weekends”*. Interestingly, this was only true more recently in the last year because she was more active during weekends and weekdays in the previous year in her data.

Table 4.3: Summary of participant insights triggered by scaffolding of the goal prompt questions. Col. 1 the participant ID. Col. 3 daily adherence (%-age of days with >1 step). Col. 4 wear-time (median hours of wear per day with >1 step). Col. 5 to Col. 8 the 4 types of insights associated with the goal prompt questions. N=21



1	2	3	4	5	6	7	8
PID	Tutorial	Daily	Hours (med)	Wkend Vs Wkdays	Work Vs Non Work	Holiday	Change Goals
P20	Y	100	18			Y	Y
P21	Y	91	13	Y		Y	Y
P6	Y	62	9				
P18	Y	44	13				
P19	Y	41	12				
P17	Y	38	11				
P7	Y	32	13	Y		Y	
P16	Y	29	N/A	Y			
P5	Y	15	14	Y			Y
P4	-	88	16	Y			
P15	-	100	17				
P10	-	100	18				
P8	-	100	16				
P9	-	100	17			Y	Y
P14	-	100	20	Y	Y		Y
P2	-	96	17				
P1	-	86	12	Y			
P13	-	70	13	Y	Y		
P11	-	60	16				
P3	-	39	14				
P12	-	31	13				
			Total	38% (8)	10% (2)	19% (4)	24% (5)
			Total	38%	10%	19%	24%

P14 was confident during the think-aloud and pre-interview that he was more active on weekend than weekdays. *"More physically active during weekends because I go hiking or something"*. However, his actual data contradicts this belief and during the think-aloud session, he did not seem to identify this. When prompted by the goal prompt panel questions, he considered this more closely and commented, *"When I see on your website, there is a lot of white [referring to weekends]. I wonder why that is"*. During post interview, when asked if he found anything surprising, he commented, *"I thought I was hitting my step goals more than I was on the weekends, it was definitely an eye opener"*. P21, commenting on doing better on weekends, *"I'm actually overshooting my goal, quite consistently"*. P7 commented on the differences between Saturday and Sundays, *"good on Saturday but not on Sundays"*.

Our participants were generally aware of their work versus non-work periods and most made observations about this during the think-aloud session when they reflected on their own data. The goal prompt questions helped 2 participants find something new, shown in Col. 6. P14 commented that he is not doing as well during work times, *"I seem to have a little more difficulty especially when I'm in the car a lot"*. P13 commented, *"Doing well when going to the office, not so well when working from home. More recently working during the week has been less good"*. During the post interview, when asked whether he learned something new, he commented, *"Going through the Goal prompts, it got me to think about things like looking at weekends versus during the week – it definitely got me thinking about different ways to divide up my time"*. This question did not apply to 2 participants (P9,P1) because they did not work or study. For example, P9 commented, *"doesn't matter because weekend and weekdays are pretty much the same to me, I'm retired"*.

During the main think-aloud exploration of *iStuckWithIt*, most participants reflected on the effect of holidays – this was triggered by visible trends in their own goal achievement. The goal prompt question (Col. 7) helped 4 participants (19%) find more. For example, P21 went back to the *iStuckWithIt* display, reviewing holiday periods more closely and commented, *"I'm mixed on holidays, there are some holidays where I do walking, there are holidays where I was in [location], ... , wasn't wearing it back there, I was wearing it sometimes but not much"*. P7 reflected on their broad wearing patterns, daily and hourly, during holidays, *"generally I forget to wear it, and I don't take it often and I fear losing it, and I don't want to take all the charging cables"*. Others focused on goal achievement consistency, for example, P9 commented, *"I walked on Christmas day. I walked on New Year's Day – so I'm quite consistent"*.

While most participants reflected upon both their activity levels and wearing behaviour in the main session exploring *iStuckWithIt*, the goal prompt questions helped 4 participants (19%) to consider changing their goal targets. For example, P21 commented, *"I guess I should lower it to something more realistic to achieve that the 10K steps per day. Maybe if I change to 8K"*. P14, after realising that he was less active on weekends commented, *"I want to make sure I hit my step goals more on the weekends, walking a little bit more"*. P9 commented, *"I should change my goal because I'm exceeding it too easily."*. Notably, P4 was confused by the wording of the prompt and failed to see the usefulness of this question. She commented, *"I don't understand why this question is here, I compare myself to myself, it doesn't matter to me"*.

Not all participants considered the goal prompts useful. For example, P16 commented, *"I think the ideas of these questions are interesting but these ones are not as useful for me. I prefer it gives me advice or record what I did"*. P16 also suggested that a prompt about how he achieved a goal each day might be useful in helping him remember. *"I think if each day when I achieve the goal, I get a question about how I achieved the goal, I think that could be useful. It might be useful if it can help me remember how I achieved the goal"*. P21 commented, *"I'm not sure how I would use it, I don't have so many goals that I have to write them down"*. He went on to suggest that these types of prompts might be better integrated into the regular weekly email that that Fitbit currently sends – he suggest this should also highlight notable items, like peaks and ask for reflection on those days then. *"I would like to be prompted as part of the weekly email if there was something interesting e.g., 28K steps days, to note things like I went hiking into the note. It would be nice to have it integrated with my weekly email to complement what I currently see"*.

To summarize, the goal prompt helped 10 participants (48%) gain new insights and 7 of these participants gained two or more insights. Broadly, our results indicate that the goal prompt scaffolding for reflection is valuable.

4.5 Discussion

To understand our long-term OLM, and reflect on their long-term behaviour and its implications, users needed to *make sense of their long-term data*. Our interface transforms that data into an OLM which highlight two key aspects. It shows a person's *apparent goal achievement*, in terms of how they did compared with their target for steps per day, active minutes or distance. However, to related this to the actual activity level, they also need to appreciate the *implications of their wearing behaviour*. This determines the completeness and meaningfulness of their data. The tutorial scaffold was intended to both support this and to enable us to study how readily people could learn about these aspects with all tutorial condition participants doing this in the context of two sets of carefully designed data, for the hypothetical users Alice and Alex. This section first considers what we learnt from the study of the tutorial scaffolding. We then discuss what our findings reveal about goal prompt scaffolding. Finally, we discuss lessons learned and insights for designing future OLMs for this class of life-long, life-wide learning for an important aspect of health.

No tutorial is needed to understand daily goal achievement trends (RQ1)

Our study results show that all participants, in either the tutorial or control group found it easy to see that blocks of missing data (grey cells) meant there was no data and that other cells indicated whether they met their goal. In the tutorial condition, participants could do this without assistance for the case of both Alice and Alex. They then went on to make a similar interpretation of their own data. The control condition participants also found this aspect of the interface intuitive and a good basis for reflection about the reasons they met their goals, commenting on aspects like holidays, injury and motivation. So, the tutorial was not needed for this case.

The tutorial did not make a difference in helping people understand wear-time (RQ1)

We observed that a minority of participants could *discover this* from the carefully crafted data for Alice and Alex. Only 56% of participants noted the low wear-time for Alice in the later months of tracking. Even fewer speculated or commented on its meaning. Moreover, only 2 participants noticed inconsistent wear-time for Alex over the long-term. Our design of the tutorial was based on providing very clear cases that should have made this concept easier to discover. For the participants who did not work this out, the experimenter explained it to them. Our observations of participants studying their own data indicate that both the tutorial and control condition participants performed similarly on this aspect. So, the tutorial was not helpful for this case and further, many participants found it difficult to appreciate the concept of wear-time. This is unfortunate since it is critical to take account of the number of hours with data to judge whether there is enough data to conclude about whether they have met their goal. There appear to be two main ways to tackle this problem. First, we may need to help people appreciate the importance of taking account of wear-time, particularly the number of hours of wear in a day. Second, the interface should make this clearer than it currently does.

Can scaffolding support insights? (RQ2)

We designed *iStuckWithIt* to present and overview model of people's goal achievement for levels of physical activity. The goal prompt scaffolding was designed to help users reflect on questions that they may not have considered when reviewing their own data. Our study showed that even for our population of existing long-term physical activity trackers who are highly educated and familiar with technology, the goal prompt scaffolding could help many of them reflect on important aspects that they had not considered. Our results showed that while *iStuckWithIt* design was useful in supporting reflection on its own, the goal prompt scaffold helped many to consider and discover insights they missed.

Adherence scaffolding design: lessons learned

In this section, we discuss lessons learned and opportunities for future user interfaces that aims to support daily adherence and wear-time for reflecting on long-term physical activity data.

First, our results suggest that a more adaptive or personalised approach is needed to teach about wear-time. This should take account of the user's actual wear-time. For people who had high wear-time (e.g., 15 hours or more per day) and consistent wear (near 100% of days have ¿0 steps), there is no need to consider wear-time when interpreting the interface's goal achievement display. In this case, future interfaces could highlight just the low wear-time days. The configuration interface could be enhanced to define suitable thresholds for this. Then the interface could filter the display to show only days that meet the threshold. For those who do *not* have high wear-time, further work is needed. However, since our long-term physical activity tracker participants had high wear-time we have limited information for this type of user.

The goal prompt scaffolding findings demonstrate the insights gained by half our participants, by considering salient aspects. This is also an opportunity for personalisation, so that the cases that deserve attention are provided as prompts. This is particularly likely to be valuable for real world use, rather than our laboratory study. It will be important in ensuring users can readily tackle the challenge of reflection, when confronted for the first time with an unfamiliar OLM interface for their long-term performance on physical activity. Opportunities for scaffolding include highlighting aspects known to be important, such as weekend versus weekdays, asking users to consider their performance and behaviour across different environments (e.g., work versus nonwork). Our findings show that there is a need to personalise such prompts based on the individual's behaviour. Prompts or questions that do not apply to the participants or do not fit the user's circumstances (e.g., work situation) or goals should not be shown.

Finally, it may be useful to send goal prompts or reflective questions via regular email messages to capture contextual and qualitative information about days of interest (e.g., what did they do or how they achieved a peak day or goal-met day). Participant comments suggest they believe capturing such data may help them remember to consider and reflect on goal prompt questions.

4.5.1 Limitations

Our study was restricted to existing FitBit users. As FitBits had been widely available for many years [5], this allowed us to recruit participants who has already collected long-term data. Also, our study is a lab study of the scaffolding designs. This may limit generalisability of our findings for wider populations of activity trackers in authentic settings. Moreover, since the iStuckWithIt user interface and study was designed for

[5] http://www.wareable.com/fitbit/fitness-tracker-sales-2015-fitbit-1169

long-term data, the usefulness of our scaffolding design on short term data was not examined. Our work paves the way for longer term field studies. Further work is still needed on scaffolding designs that help people understand the impact of data incompleteness in short term physical activity tracker data.

4.6 Conclusion

In this study, we explored how to help people reflect on their long-term physical activity goal achievement. We extended previous work by exploring two forms of scaffolding (tutorial, goal prompt) and reported on a lab study of 21 existing long-term physical activity tracker's experiences. The tutorial scaffolding results reveal that missing or incompleteness in data on a day-to-day basis (daily adherence) is intuitive and well understood. However, for wear-time (hours of wear per day), it may be more appropriate to provide prompts and alerts based on automatic detection of features such as low wear-time on certain days, weeks or highlighting inconsistencies over time. In addition, our prompt scaffolding (reflective questionnaire) proved effective support for reflection, resulting in new insights. We go beyond previous OLM work, particularly in the focus on a lifelong, life-wide learning goal associated with long-term physical activity goal achievement. Our findings provide design insights about these two scaffolding approaches apply, with recommendations on future work and potential roles for reflection scaffolding and its personalisation.

Chapter 5

Defining Adherence: Making Sense of Physical Activity Data

Preamble

> Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., and Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):37

This work was published in the ACM journal on Interactive, Mobile, Wearable and Ubiquitous Technologies in 2018. It reported on contributions described in section 1.2.4 and 1.2.4.

Abstract

Increasingly, people are collecting detailed personal activity data from commercial trackers. Such data should be able to give important insights about their activity levels. However, people do not wear or carry tracking devices all day, every day and this means that tracker data is typically *incomplete*. This paper aims to provide a systematic way to take account of this incompleteness, by defining *adherence*, a measure of data completeness, based on how much people wore their tracker. We show the impact of different adherence definitions on 12 diverse datasets, for 753 users, with over 77,000 days with data, interspersed with over 73,000 days without data. For example, in one data set, one adherence measure gives an average step count of 6,952 where another gives 9,423. Our results show the importance of adherence when *analysing* and *reporting* activity tracker data. We provide guidelines for defining adherence, analysing its impact and reporting it along with the results of the tracker data analysis. Our key contribution is the foundation for analysis of physical activity data, to take account of data incompleteness.

5.1 Introduction

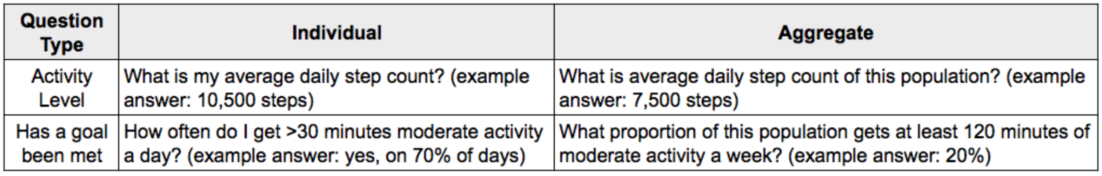
Millions of people track their physical activity. Eighteen percent of US adult consumers own a wearable fitness tracker, with the majority using their device often [Albert, 2017]. Over 22 million wearable devices were shipped in the second quarter of 2016 alone [IDC, 2017]. Beyond this, smart phones and watches are increasingly making it possible for more people to track their physical activity.

This means that people are building up vast collections of data about their physical activity. Such data should be valuable for the *individual* who wants to understand their own long-term activity. The *aggregated data* can provide a low cost way to collect data about various populations. Table 5.1 shows examples of the types of questions that such data have the potential to answer. The first row illustrates questions about *activity level*. For example, an individual may want to know how many steps a day they average; an example answer is 10,500 steps. Answering such questions can give the benefits for reflection, self-monitoring and planning as documented in personal informatics research [Consolvo et al., 2014, Meyer et al., 2016a, Gouveia et al., 2015, Epstein, 2015, Fritz et al., 2014, Li et al., 2012, Rooksby et al., 2014]. *Aggregate* *analyses* can provide a corresponding average daily step count, such as the 7,500 for that population. A large body of health literature has examined such questions, for example, to inform recommendations about levels of physical activity for good health [Tudor-Locke et al., 2011, Haskell et al., 2007]. The second type of question asks whether a target *goal has been met* [Tudor-Locke et al., 2011, Tucker et al., 2011]. For example, did I meet my goal of >30 active minutes a day.

A key challenge for interpreting physical activity data is that it is typically incomplete [Meyer et al., 2017, Epstein et al., 2016a, Tang and Kay, 2016, Tang and Kay, 2017]. Many factors can contribute to gaps in the data, such as forgetting to wear the device, device loss or changes in motivation to track [Epstein et al., 2016b, Meyer et al., 2017, Tang and Kay, 2016, Consolvo et al., 2014, Bentley et al., 2013]. To answer questions like those in Table 5.1, it is critical to account for this incompleteness. To see why this is so, consider the first question in the table – determining a person's average step count. Consider the case of Alice, who wears her tracker all day, every day; a reliable answer can be calculated as a simple average over each day's step counts. But consider another person, Bob, who wears his tracker only on 60% of days – but on those days he wears it at all, he wears all day. His average daily step count should be based on *just those days where he has data* (the total step count divided by the days with data). We also need to consider the impact of the wear time within a day. Consider another person, Carol, who wears her tracker every day but only *in the morning on weekends* and *all day on weekdays*. Now a meaningful answer is more complex to determine – it needs to account for the incompleteness of her weekend data.

This paper aims to provide foundations for a systematic process to take account of the incompleteness of personal sensor data for physical activity when answering questions like those in Table 5.1. We introduce ***adherence***, a notion that reflects the

Table 5.1: Examples of important questions long-term physical activity data can answer, at the level of an individual or aggregate, population levels.



Question Type	Individual	Aggregate
Activity Level	What is my average daily step count? (example answer: 10,500 steps)	What is average daily step count of this population? (example answer: 7,500 steps)
Has a goal been met	How often do I get >30 minutes moderate activity a day? (example answer: yes, on 70% of days)	What proportion of this population gets at least 120 minutes of moderate activity a week? (example answer: 20%)

fact that an activity tracker should give accurate answers to questions about activity for people like Alice who has 100% adherence, wearing her tracker all day, every day. We aim to establish adherence measures to account for people with less than 100% adherence, be this like Bob, Carol or the myriad of other wearing possibilities.

While previous work has studied wearing behaviour [Bentley et al., 2013, Consolvo et al., 2014, Epstein et al., 2016a, Epstein et al., 2016b, Meyer et al., 2017, Tang and Kay, 2016, Tang and Kay, 2017], there has been no research on how to systematically tackle the analysis and reporting of that data to account for its incompleteness. This is important if people are to *trust* the information that applications report on physical activity, claiming, or appear to claim, that ubicomp sensor data gives objective truths of one's activity. Fogg [Fogg, 2003] warns that when systems produce questionable data, people are less likely to trust them and so they are less useful as a behaviour change tool. Bentley et al [Bentley et al., 2013] found that incomplete data led to a loss of trust in their tool. Consolvo et al ([Consolvo et al., 2014], page 234) also reported this when missing data affected self-monitoring feedback.

To address these challenges, our work aims to provide systematic foundations for analysis and reporting physical activity data, accounting for data incompleteness. To do this we tackle the following research questions:

RQ1 What is the impact of different adherence measures on data ignored?
RQ2 How can we account for adherence for Activity-level questions?
RQ3 How can we account for adherence for Goal-met questions?

We explored these questions by analysing the impact of different adherence definitions on 12 datasets, with a total of 753 physical activity tracker users, who had more than 77,000 days with data, interspersed with over 73,000 days without data. We analysed them with 4 adherence measures that have been reported in previous literature.

- >0 steps: the least stringent answers activity questions using data from any day that has any data [Epstein et al., 2016b, Meyer et al., 2016b, Meyer et al., 2017, Tang and Kay, 2016, Tang and Kay, 2017],

- >500 uses only days with more than 500 steps [Meyer et al., 2016b, Meyer et al., 2017];

- >10 hours – uses only days with at least 10 different hours with data [Migueles et al., 2017, Tang and Kay, 2016, Cadmus-Bertram et al., 2015b];

- 3-a-day one requires data within 3 time periods of the day [Meyer et al., 2016b, Meyer et al., 2017].

As this is the first work to establish a systematic way to account for adherence, we chose this core set of research questions, 12 very diverse datasets and these 4 adherence measures from the literature.

The next two sections introduce key terminology and related work. Then the core of the paper: the study design, results and their discussion.

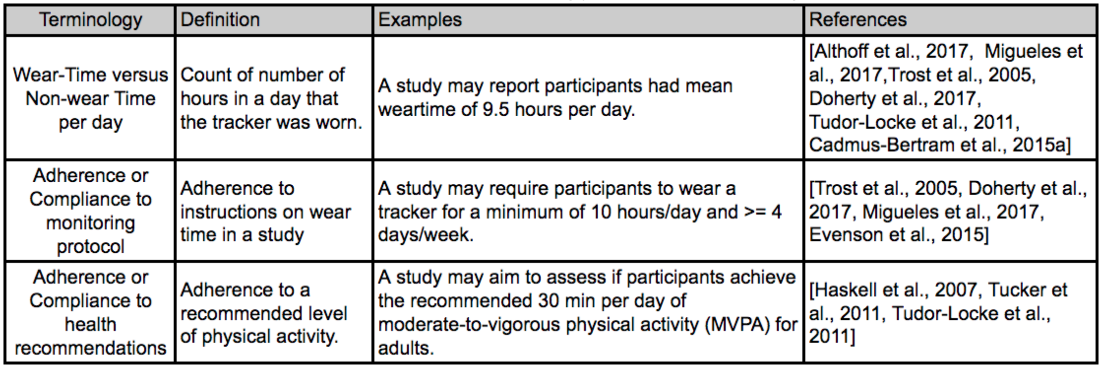
5.2 Definitions of Adherence

This section introduces key terms for defining adherence. It begins with ways to describe tracker wear-time. It then defines a *valid day*, one with data of sufficient quality that it is meaningful to include in analyses, along with criteria for assessing the validity of a day and ways to report adherence. It concludes with a review of adherence that goes beyond a single day, to describe adherence for one week and for longer periods.

5.2.1 Adherence & Wear-time

Table 5.2 introduces terms that have been used to describe wear-time of trackers. The first row shows how *wear-time* has been expressed as the number of hours of wear-per-day. It also shows non-wear time. For devices like the Fitbit, it is difficult to distinguish inactivity (e.g., 0 or few steps) from non-wearing. However, Migueles et al [Migueles et al., 2017], reviewing use of accelerometer data in physical activity studies, reports that 20 consecutive minutes with very low accelerometer activity reliably indicates non-wear time for adults. The second row shows one common use of the term, adherence, to mean that study participants wore their tracker for the required number of hours per day and days per week. This is used to determine whether each day's data, or a person's full dataset, has good enough adherence to be a reliable measure of their actual activity level. Doherty et al [Doherty et al., 2017] reported the impact of various factors, including age, sex, day, time of day, and season on this adherence measure. We adopted this meaning for adherence. However, our work aims to go beyond analysis of study data where participants are recruited to wear a tracker as instructed. We want to be able to meaningfully analyse the vast collections of activity data that that people are building up. For this, we need to refine the notion of adherence to provide a conceptual framework and a systematic process for ubicomp researchers, and others, to meaningfully interpret such data.

Table 5.2: Adherence terminology used in existing literature.



Terminology	Definition	Examples	References
Wear-Time versus Non-wear Time per day	Count of number of hours in a day that the tracker was worn.	A study may report participants had mean weartime of 9.5 hours per day.	[Althoff et al., 2017, Migueles et al., 2017,Trost et al., 2005, Doherty et al., 2017, Tudor-Locke et al., 2011, Cadmus-Bertram et al., 2015a]
Adherence or Compliance to monitoring protocol	Adherence to instructions on wear time in a study	A study may require participants to wear a tracker for a minimum of 10 hours/day and >= 4 days/week.	[Trost et al., 2005, Doherty et al., 2017, Migueles et al., 2017, Evenson et al., 2015]
Adherence or Compliance to health recommendations	Adherence to a recommended level of physical activity.	A study may aim to assess if participants achieve the recommended 30 min per day of moderate-to-vigorous physical activity (MVPA) for adults.	[Haskell et al., 2007, Tucker et al., 2011, Tudor-Locke et al., 2011]

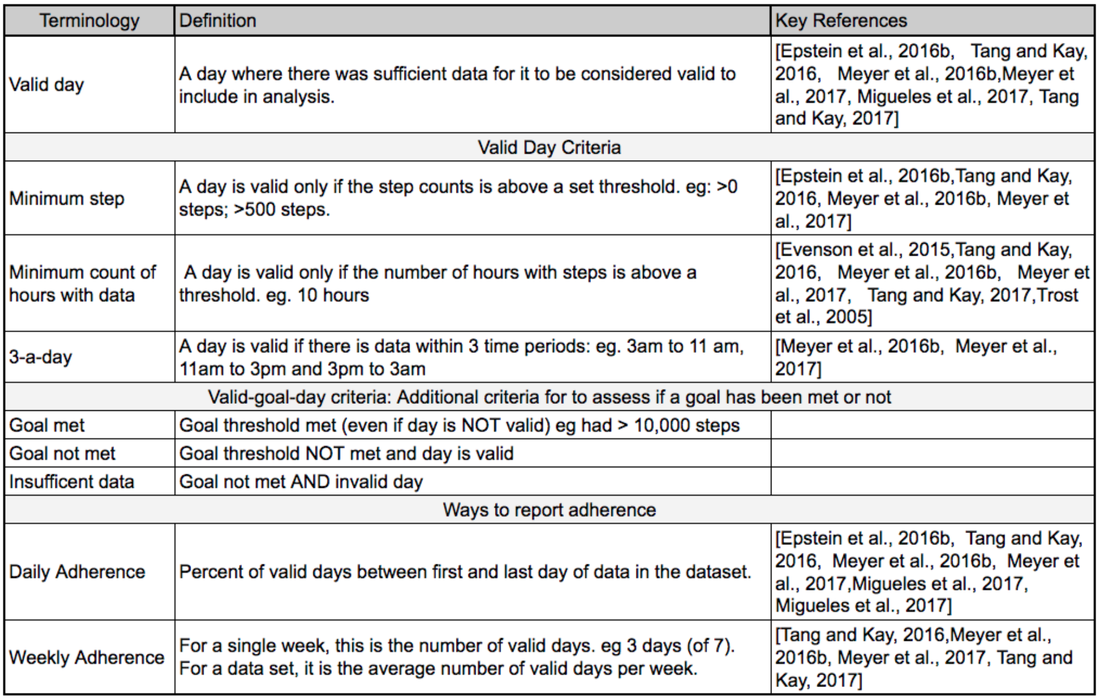
The last row, *adherence to physical activity recommendations* (also called *compliance*), describes how well a person or sample population meets a recommended level of physical activity. Adherence to a recommendation is of enormous importance [Haskell et al., 2007] and the use of trackers to obtain objective measures of activity levels is of intense interest in health literature [Evenson et al., 2015, Migueles et al., 2017, Doherty et al., 2017, Tucker et al., 2011, Tudor-Locke et al., 2011].

5.2.2 Adherence measures

Table 5.3 introduces terms for describing and defining adherence. The key notion is the *Valid day*, a day with enough data to justify including it in analyses. Any data from non-valid days are ignored. The next 3 rows describe valid day criteria that have been used in previous work, listed in the last column. The simplest is based on a step threshold. When the threshold is >0 steps [Epstein et al., 2016b, Tang and Kay, 2017, Tang and Kay, 2016] any day with any step data is considered valid. The >500 steps, used in [Meyer et al., 2016a, Meyer et al., 2017] requires at least 500 steps for a valid day.

The second and third rows are ways to determine if there is enough data *through the day* to make that day valid. The 10-hours criterion means that there are at least 10 hours in the day that each have at least one step [Tang and Kay, 2017,

Table 5.3: Adherence terms and definitions used in this study.



Terminology	Definition	Key References
Valid day	A day where there was sufficient data for it to be considered valid to include in analysis.	[Epstein et al., 2016b, Tang and Kay, 2016, Meyer et al., 2016b,Meyer et al., 2017, Migueles et al., 2017, Tang and Kay, 2017]
Valid Day Criteria		
Minimum step	A day is valid only if the step counts is above a set threshold. eg: >0 steps; >500 steps.	[Epstein et al., 2016b,Tang and Kay, 2016, Meyer et al., 2016b, Meyer et al., 2017]
Minimum count of hours with data	A day is valid only if the number of hours with steps is above a threshold. eg. 10 hours	[Evenson et al., 2015,Tang and Kay, 2016, Meyer et al., 2016b, Meyer et al., 2017, Tang and Kay, 2017,Trost et al., 2005]
3-a-day	A day is valid if there is data within 3 time periods: eg. 3am to 11 am, 11am to 3pm and 3pm to 3am	[Meyer et al., 2016b, Meyer et al., 2017]
Valid-goal-day criteria: Additional criteria for to assess if a goal has been met or not		
Goal met	Goal threshold met (even if day is NOT valid) eg had > 10,000 steps	
Goal not met	Goal threshold NOT met and day is valid	
Insufficent data	Goal not met AND invalid day	
Ways to report adherence		
Daily Adherence	Percent of valid days between first and last day of data in the dataset.	[Epstein et al., 2016b, Tang and Kay, 2016, Meyer et al., 2016b, Meyer et al., 2017,Migueles et al., 2017, Migueles et al., 2017]
Weekly Adherence	For a single week, this is the number of valid days. eg 3 days (of 7). For a data set, it is the average number of valid days per week.	[Tang and Kay, 2016,Meyer et al., 2016b, Meyer et al., 2017, Tang and Kay, 2017]

Trost et al., 2005, Migueles et al., 2017, Evenson et al., 2015]. The 3-a-day measure, in [Meyer et al., 2017] is another way to ensure data through the day, this time requiring data in each of three parts of a day.

The next part of the table show the cases to consider when answering Goal-met questions. Activity level questions need a minimum wear-time for a valid day, but this may not be needed for a ***Valid-Goal-Day***. For example, with a goal of 10,000 steps and a 10-hours valid day threshold, a valid-goal-day occurs whenever the goal is met:

- *Goal met (valid day)*: i.e. reached 10,000 steps, with 10 hours wear-time,
- *Goal met (non-valid day)*: i.e. reached 10,000 steps, with 3 hours wear-time.

If the goal is not met, the day is only valid if the 10-hour valid day threshold is met:

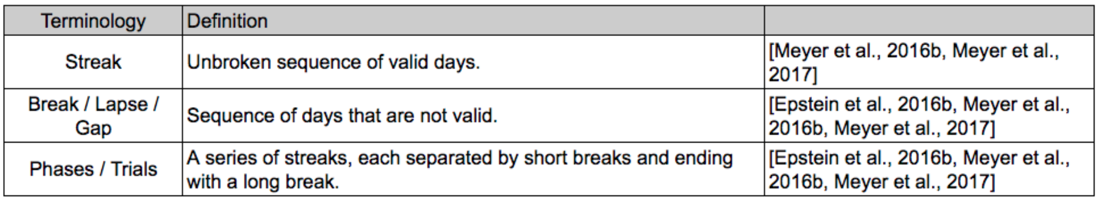
- *Goal not met + valid day*: e.g., reached 3,000 steps, with 10 hours wear-time,

Then the only days are invalid if there is *insufficient data* to make the day valid if neither the threshold is met, nor the goal e.g., reached 3,000 steps with 4 hours wear-time.

The last two rows introduce terms to use when reporting adherence. **Daily** adherence refers to the percentage of valid days in a dataset, as a description of completeness in terms of valid days. Daily adherence can be calculated for an individual or a population.

We now consider adherence beyond a single day. **Weekly** adherence measures the average number of valid days per week (only calculated during weeks where there is at least one valid day). For example, suppose a person has 50% daily adherence (i.e., having 50% of days valid) and 7 days-per-week weekly adherence. This corresponds to a person who had only 50% of their days valid but this held over every day of the weeks with data. Table 5.4 summarises other terms that have been used to describe patterns of adherence. A **streak** is an unbroken sequence of valid days. In contrast, a **break, lapse or gap** describes sequences of days that are not valid days. **Phases or trials** describe a series of streaks separated by short breaks and ending with a long break.

Table 5.4: Measures for long-term adherence patterns.



Terminology	Definition	
Streak	Unbroken sequence of valid days.	[Meyer et al., 2016b, Meyer et al., 2017]
Break / Lapse / Gap	Sequence of days that are not valid.	[Epstein et al., 2016b, Meyer et al., 2016b, Meyer et al., 2017]
Phases / Trials	A series of streaks, each separated by short breaks and ending with a long break.	[Epstein et al., 2016b, Meyer et al., 2016b, Meyer et al., 2017]

In the next section, we review the large body of work reporting physical activity tracker adherence and existing methods to analyse and address incompleteness in physical activity tracker data.

5.3 Related Work

In this section, we first review research on activity tracker wearing behaviour and how this impact data completeness. We then review work that highlights why this is important for ensuring user trust and the ability to reflect on their data. Finally, we review reported methods used to deal with data incompleteness. We describe this work using the term, adherence, as just defined, although these terms were not used by the authors.

5.3.1 Studies of Wearing Behaviour

Previous work has explicitly studied wearing behaviour and patterns. It indicates wide differences in wearing behaviour, and associated with these diverse levels of data completeness [Meyer et al., 2017, Tang and Kay, 2016, Epstein et al., 2016b]. Epstein et al [Epstein et al., 2016b] identified three distinctive groups of tracker users: 1. short term, 2. intermittent and 3. long and consistent. Clearly the intermittent users have incomplete data. Meyer et al [Meyer et al., 2016b] also reported wide differences in daily adherence, 20% to 100% of days being valid. At the same time, there are reports of some users who do sustain high adherence over longer periods, months and even years [Cadmus-Bertram et al., 2015a, Meyer et al., 2017, Epstein et al., 2016b, Tang and Kay, 2017]. Some work has studied the factors affecting wear-time, including age, gender and environment [Doherty et al., 2017, Althoff et al., 2017]; day of week [Doherty et al., 2017, Meyer et al., 2017, Tang and Kay, 2016]; time of day [Meyer et al., 2017, Doherty et al., 2017]; and the efforts demanded by some tracking devices (e.g., battery life, water resistance) [Tang and Kay, 2017, Consolvo et al., 2014, Fritz et al., 2014, Epstein et al., 2016a]. This small but growing body of work highlights that there are diverse levels and patterns of wearing behaviour and so diverse levels of data completeness.

5.3.2 Importance of accounting for incomplete data

This section reviews the work that shows the importance of dealing with the incompleteness of physical activity data for personal health and well-being applications. Consolvo et al (page 211 [Consolvo et al., 2014]) identified key design challenges for applications intended to encourage health and well-being. One of these relates to the *perceived accuracy* of trackers. For example, with their trackers, some users were disappointed that their tracker did not measure many activities such as vigorous gardening. They highlight the importance of missing data for self-monitoring feedback especially when graphs are used (page 234 [Consolvo et al., 2014]). Such displays tend to overlook missing or incorrect data that can lead to incorrect presentation of trends and patterns. In the Health Mashups system, Bentley et al [Bentley et al., 2013] integrated data from multiple sources to present health behaviour patterns, well-being data and context to promote behaviour change. One of the major challenges they faced was the incompleteness in their source data (*sparseness*) which resulted in inaccuracy in their recommendations. Several users noticed contradicting information and that led to mistrust of the results. They link this result to Fogg's [Fogg, 2003] analysis of persuasive systems, warning that systems producing questionable data are less likely to be trusted and thus they are less useful as a behaviour change tool.

Consolvo et al [Consolvo et al., 2014] suggested that presenting uncertainty can help make the self-monitoring data appear more accurate and precise. They argued the need to explore how uncertainty can be presented or managed. This view is supported by Kay et al [Kay et al., 2016] in a study of uncertainty presentations in transport schedules. Participants reported that uncertainty information helped them make better decisions and alleviate anxiety when the app information did not match their knowledge. In a study of an interface that embeds daily and hourly adherence (wear-time) information in a calendar visualisation, [Tang and Kay, 2017] this adherence information helped users reflect on their long-term activity, and to link this with their knowledge about the context. Some participants also reflected also on their wearing behaviour as well as factors affecting it. These results indicate the importance of accounting for and presenting information about incompleteness or uncertainty.

To summarize, adherence can impact the *perceived accuracy* of activity tracker data and there is evidence that presenting this may improve confidence and trust in the application as well as to support reflection.

5.3.3 Adherence requirements in studies of physical activity

Research on physical activity needs to be based on sufficient data that is of sufficient quality. A body of work has examined how many valid days within a monitoring period are required to provide a confident estimate of behaviour [Tudor-locke, 2016]. For example, systematic reviews of accelerometer based physical activity assessments [Trost et al., 2005, Migueles et al., 2017] have reported a range of criteria used, with a recommendation for at least 4 days in a 7 day monitoring period. However, they do not offer specific guidelines for longer monitoring periods except to warn that while increasing requirements for *valid-day* or valid-week can improve reliability of data, it also results in greater sample loss [Migueles et al., 2017]. There has been some work on longer duration studies that used statistical methods to estimate missing values. For example, Tudor-locke et al [Tudor-Locke et al., 2004] studied 23 participants over 1 year and used the Missing Value Analysis EM function in SPSS to estimate missing values. Similarly, a yearlong study involving 37 males and 44 females [Togo et al., 2008] also used an estimation method based on linear interpolation. Notably, both studies had very high adherence, 98% and 95% were valid over the 1 year study period. These interpolation methods presume the missing days are like the ones in the dataset, as assumption that is likely to be less reliable in datasets with lower adherence levels.

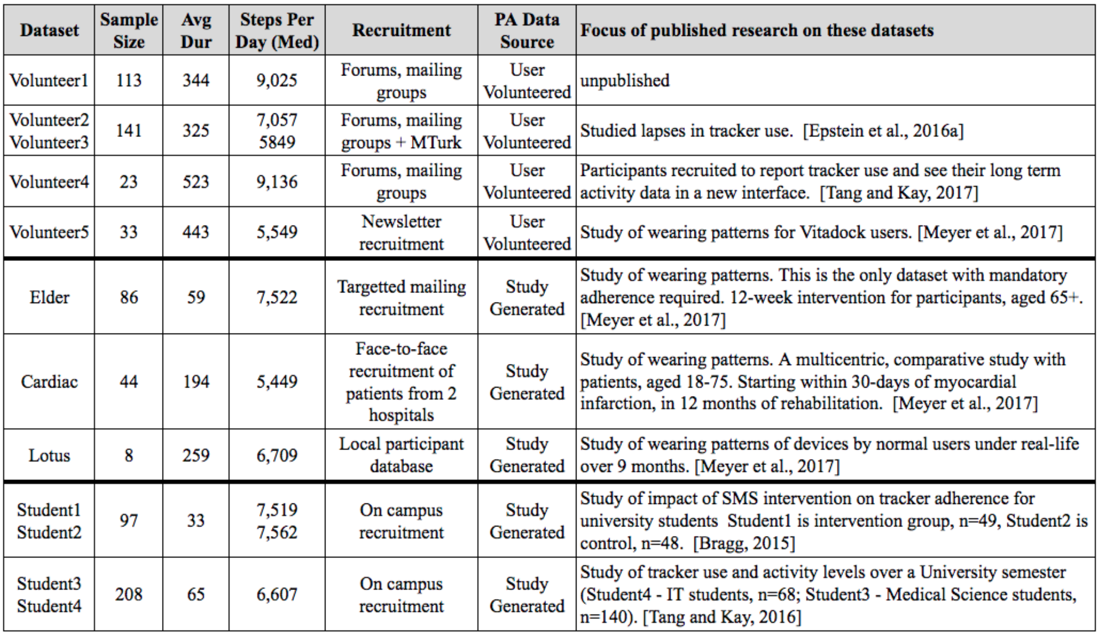
5.3.4 Summary

Existing literature highlights that people have varying wearing behaviours, providing diverse levels and patterns of completeness. This poses important problems for ensuring trust needed for people to effectively reflect on their data as a foundation for behaviour change. There has been limited work on how to address the incompleteness that can be expected in many datasets. This is the challenge that our work aims to address.

5.4 Study Design

Our study design has three elements: a suitable collection of datasets; a set of adherence definitions to explore; and a sequence of experiments to perform to gain insights into our research questions. The design process began when the Sydney University authors were analysing several of their datasets and began to appreciate the need to for a systematic approach to taking account of tracker adherence. The team initiated the collaboration with the other authors, Meyer and Epstein, to discuss their insights, based on their work on wearing behaviours and patterns, including streaks, breaks, lapses, phases and trials as just described. The new team then established the design for this study.

Table 5.5: The 12 datasets from 9 studies of various lengths and population size. The first column is the identifier we use to describe the dataset. Next is the sample size and average duration in days, the average step count (using only days with >0 steps) and then the recruitment methods. The data source column distinguishes *volunteers* datasets (the first block), from the remainder, being *other study-generated* datasets.



Dataset	Sample Size	Avg Dur	Steps Per Day (Med)	Recruitment	PA Data Source	Focus of published research on these datasets
Volunteer1	113	344	9,025	Forums, mailing groups	User Volunteered	unpublished
Volunteer2 Volunteer3	141	325	7,057 5849	Forums, mailing groups + MTurk	User Volunteered	Studied lapses in tracker use. [Epstein et al., 2016a]
Volunteer4	23	523	9,136	Forums, mailing groups	User Volunteered	Participants recruited to report tracker use and see their long term activity data in a new interface. [Tang and Kay, 2017]
Volunteer5	33	443	5,549	Newsletter recruitment	User Volunteered	Study of wearing patterns for Vitadock users. [Meyer et al., 2017]
Elder	86	59	7,522	Targetted mailing recruitment	Study Generated	Study of wearing patterns. This is the only dataset with mandatory adherence required. 12-week intervention for participants, aged 65+. [Meyer et al., 2017]
Cardiac	44	194	5,449	Face-to-face recruitment of patients from 2 hospitals	Study Generated	Study of wearing patterns. A multicentric, comparative study with patients, aged 18-75. Starting within 30-days of myocardial infarction, in 12 months of rehabilitation. [Meyer et al., 2017]
Lotus	8	259	6,709	Local participant database	Study Generated	Study of wearing patterns of devices by normal users under real-life over 9 months. [Meyer et al., 2017]
Student1 Student2	97	33	7,519 7,562	On campus recruitment	Study Generated	Study of impact of SMS intervention on tracker adherence for university students Student1 is intervention group, n=49, Student2 is control, n=48. [Bragg, 2015]
Student3 Student4	208	65	6,607	On campus recruitment	Study Generated	Study of tracker use and activity levels over a University semester (Student4 - IT students, n=68; Student3 - Medical Science students, n=140). [Tang and Kay, 2016]

5.4.1 Datasets

The team collected the 12 datasets presented in Table 5.5. This collection is diverse on many dimensions as we now describe. The first column shows the name we use to refer to the dataset. The second is the number of people in the dataset (col 2) and then is the average duration in days (col 3), from the first to the last day with data. The forth column is the median steps per day calculated using the >0 steps threshold. The next two columns (col 5 and 6) indicate the sources of the data in terms of the means used to recruit participants and whether the data was *volunteers* or *other study-generated*. The last column overviews the ways the datasets have been used in previous published work or that it has not been published. The table organised groups as *volunteers* datasets first. The remaining datasets are all *other study-generated* because participants were recruited as part of another study and in these, participants were provided with a tracker.

The first 5 datasets are *volunteers*. In these, people who were already tracking were recruited to volunteer their data. *Volunteer1* consists of 113 Fitbit users who tracked from 18 to 731 days (average 344), with 73% having tracked for >=6 months. The *Volunteer2* and *Volunteer3* Fitbit datasets were used to study gaps and lapses in activity tracker use [Epstein et al., 2016a]. These had different recruitment methods: of the 141 users, the 67 of *Volunteer2* were recruited in a similar way to *Volunteer1* and the other 74 via Amazon Mechanical Turk (a popular crowd-sourcing website). The two groups had different Fitbit use patterns, with participants from the snowball recruitment wearing their Fitbits more and walking more each day than those recruited via Amazon Mechanical Turk. *Volunteer4* recruited long-term Fitbit trackers via forums and email, for a study that sought to understand how they already used their long-term activity data and then to study their use of a calendar-based interface showing their full record of activity and wearing behaviour [Tang and Kay, 2017]. All but 2 had >6 months of data. The last in this group, *Volunteer5*, also a volunteered dataset, involves a different device, the VitaDock [Meyer et al., 2017].

All the remaining datasets are *other study-generated*, meaning that the data was collected as part of another study for which participants were recruited to answer a question unrelated to studying wearing behaviour. The middle block of three datasets and these, along with *Volunteer5* were analysed by Meyer et al [Meyer et al., 2017] to gain understanding of wearing patterns and how to describe them. These three studies used various trackers, including various versions of Fitbit and the Medisana ViFit tracker. The first, *Elder*, is distinctive in that it is the *only dataset in our collection where tracker use was mandatory*, as is typical in medical studies. This means that this dataset only has participants who had recorded tracker data. This dataset is also distinctive as it recruited an older population, aged 65-75. *Cardiac* users were recruited within 30 days of a myocardial infarction as part of a 12-month rehabilitation program. *Lotus* is a very small longitudinal observational study with no explicit intervention. This study's population is closer to the self motivated Fitbit users from *Volunteer1*, *Volunteer2*, *Volunteer3* and *Volunteer4*. These 3 studies used various activity tracking devices including various versions of Fitbit and the Medisana ViFit tracker.

The next 4 data sets came from studies of university students who were lent a Fitbit Zip. *Student2* and *Student1* involved Medical Science students, recruited in a tutorial class and split into control group (*Student2*) and an intervention group (*Student1*) to assess the impact on wear-time from a weekly SMS (text message) on Fridays, reminding the experimental group to wear their tracker. [Bragg, 2015]. *Student3* and *Student4* datasets were from an observational study to learn about physical activity levels of undergraduate students [Tang and Kay, 2016]. These students were also recruited in a tutorial class.

To summarize, our datasets are diverse in terms of all the dimensions summarised in Table 5.5 as well as the details above. This makes it a rich collection for exploring our research questions.

5.4.2 Thresholds

Table 5.3 introduced several definitions and background for defining the thresholds for a valid day. As this is the first systematic analysis of adherence over a diverse collection of datasets, we restricted our analyses to just four carefully chosen valid day thresholds that have been used in previous research on wearing behaviours and in analysing activity tracking data:

- >0 steps as this is the simplest and least restrictive – used in [Epstein et al., 2016a, Tang and Kay, 2016]
- >= 500 steps as another simple step criterion – used by [Meyer et al., 2016b, Meyer et al., 2017]
- 3-a-day, a measure of wear through the day [Meyer et al., 2016b, Meyer et al., 2017]
- >10 hours, the most stringent measure, requiring at least 10 hours, each with at least 1 step. [Tang and Kay, 2016, Tang and Kay, 2017, Migueles et al., 2017]

The 10-hours adherence threshold has been used in health literature [Migueles et al., 2017, Tudor-Locke et al., 2008, Matthews et al., 2008, Trost et al., 2005] - although lesser ones have also been used (e.g. 8 hours [Konstabel et al., 2014]).

5.4.3 Analysis conducted

We conducted the experiments to compare:

1. %-age of *days* excluded using each of the 4 thresholds above;
2. %-age of *people* whose median wear-time (hours per day) exceeds our >10 hours threshold, comparing this with the proportion whose wear-time was 6-9 hours and <6 hours.

3. how the 4 thresholds affect calculated *activity levels*.

In each of these broad categories, we explored the impact of the across the datasets and tested for significant differences.

In our analysis, wear-time (hours per day) is calculated as the count of hours where at least 1 step was recorded. This approximation is needed because the trackers for our datasets do not distinguish a sedentary person wearing the tracker from a tracker that is not worn.

While our method is not exact for accounting non-wear, it serves as an estimate for comparison at a population level.

5.5 Results

We now present the results of our analysis. The presentation is organised around our three research questions. First, we consider the impact of different adherence definitions on the data ignore (RQ1). Then we show how these definitions impact on results. In the discussion, we consider how these results point to ways to account for these results (RQ2 and 3).

5.5.1 Impact of thresholds on valid days and valid weeks (RQ1)

At a high level, we would expect that the more restrictive a threshold is, the more data would be excluded from analysis of physical activity data. However, it is not obvious what differences will follow from different adherence definitions used to analyse datasets.

Differences in the Days ignored for different adherence measures

Figure 5.1 compares the %-age of days of data that are ignored for dataset. It takes the >0 steps measure as a baseline, since it is the least stringent measure. It shows how the other three measures compare against it. This and subsequent graphs order the datasets as in Table 5.5. This groups them according to the broad characterisation of the datasets in terms of whether they were volunteered or part of a study that influenced participants to track. We include the sample size in the labels so the reader can see where the small size of a dataset may help explain the results.

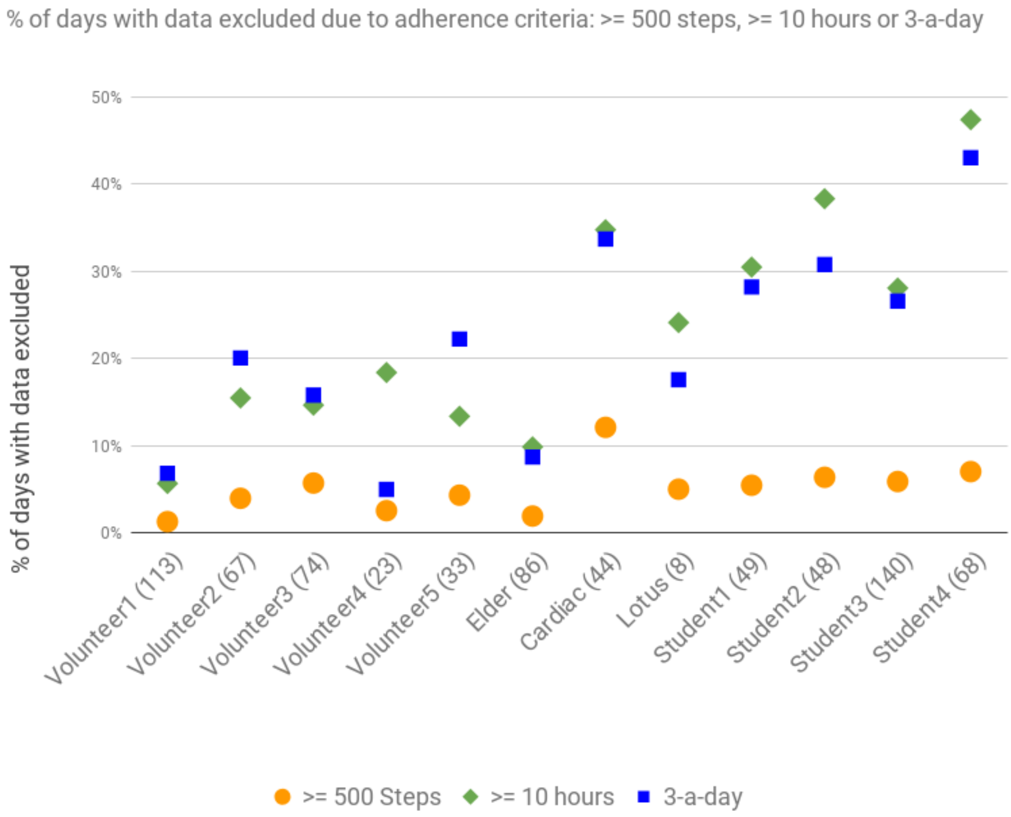


Figure 5.1: Comparing %-age of days discarded, with valid day thresholds: >=500 steps, >=10-hours and 3-a-day, against >0 steps − days with no data

First consider the >= 500 minimum step criterion (yellow circles). Overall, the graph shows that this criterion discards between 5% and 10% of days that are valid on the >0 steps criterion (mean: 5.1%, SD: 2.8%, 95% CI: ±1.8%). These are days with exceedingly modest use of the tracker (i.e., between 1 and 500); if a dataset has many of these, the reasons for this may deserve exploration to understand why people would often make so little use of a tracker or if it indicates problems with the device (such as problems with device calibration for people with specific mobility problems, and using a walking frame).

We now consider the *through the day* thresholds (green diamond for >= 10 hours and the blue square for 3-a-day). There are three striking trends here. First, the levels of data loss now have a far wider range, from 6% (*Volunteer1*) to 47% (*Student4*). Secondly, the data loss is always higher than the >=500 step threshold; these differences are significant (one-way ANOVA F(2,33)=5.64, p<0.001). Thirdly, both *through the day* thresholds are strikingly similar to each other for most datasets, also reflected in the mean %-age of days of data discarded showing no significant differences.

- mean: 21.5%, SD: 11.7%, 95% CI: ±7.4% - 3-a-day threshold
- mean: 23.4%, SD: 12.7%, 95% CI: ±8.1% - 10-hour threshold

One clear outlier is *Volunteer4*, the only case where 3-a-day is around 5% and much closer to the >= 500 steps and the 10 hours threshold is almost 20%. shown in Figure 5.1. This may be due to the small sample size (N=23) or variations at the individual level. As we would expect, these results show that the *through the day* thresholds, being stricter, may discard far more data. However, the differences varied widely across the datasets. The figure indicates much smaller differences for the *Volunteers* datasets, compared with the other datasets (Students plus Others - *Elder*,*Lotus*,*Cardiac*) and the difference is significant (2 sample t-test, t_{10}=3, p=0.01, 95% CI: ±12.6%):

- mean data loss: 13.5%, SD: 4.8%, 95% CI: ±5.9% for the *Volunteers* datasets;
- mean data loss: 30.4%, SD: 11.8%, 95% CI: ±10.9% for the other datasets (Students plus Others - *Elder*,*Lotus*,*Cardiac*).

The example in Figure 5.2 illustrates that this issue is also relevant when answering Goal-met questions. In this example, using a valid-day threshold of >10 hours, we show the percent of days that would be excluded (or considered insufficient data days)

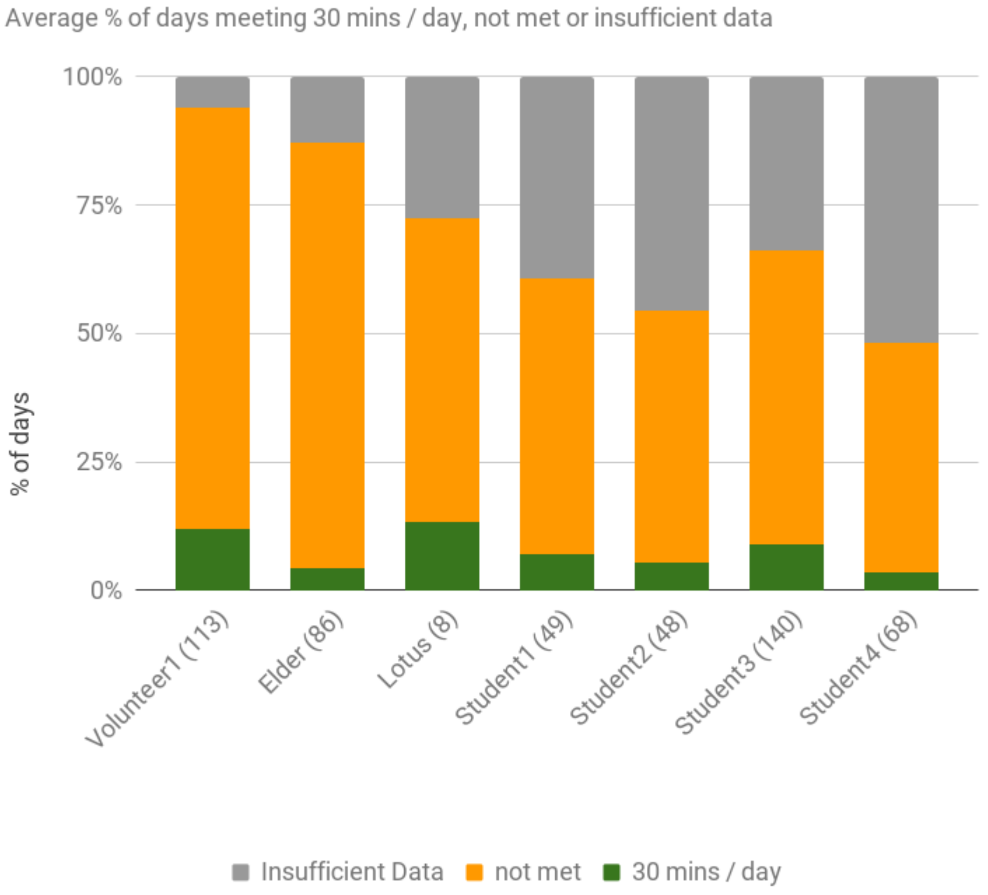


Figure 5.2: An example of a Goal-met report showing percentage of days that met 30 active minutes, not met or had insufficient data. Note: using 10 hours valid threshold to determine insufficient data days. Note: datasets *Volunteer2*, *Volunteer3*, *Volunteer4*, *Cardiac* and *Volunteer5* are not included due to lack of per minute data. N of each sample included in X-axis label.

for each of our datasets. The green bars represent percentage of days where participants meet the 30 active minutes goal (i.e., defined as goal-met in Table 5.3). The yellow bars represent percentage of days where participants did not meet goal and recorded 10 hours or more data (i.e., goal-not-met). The grey bars represent the percentage of days where participants did not meet the goal but also did not record 10 hours of data (i.e., insufficient data). The figure shows very wide differences in the percentage of insufficient data days across different datasets from 6% for *Volunteer1* to over 52% for *Student4*.

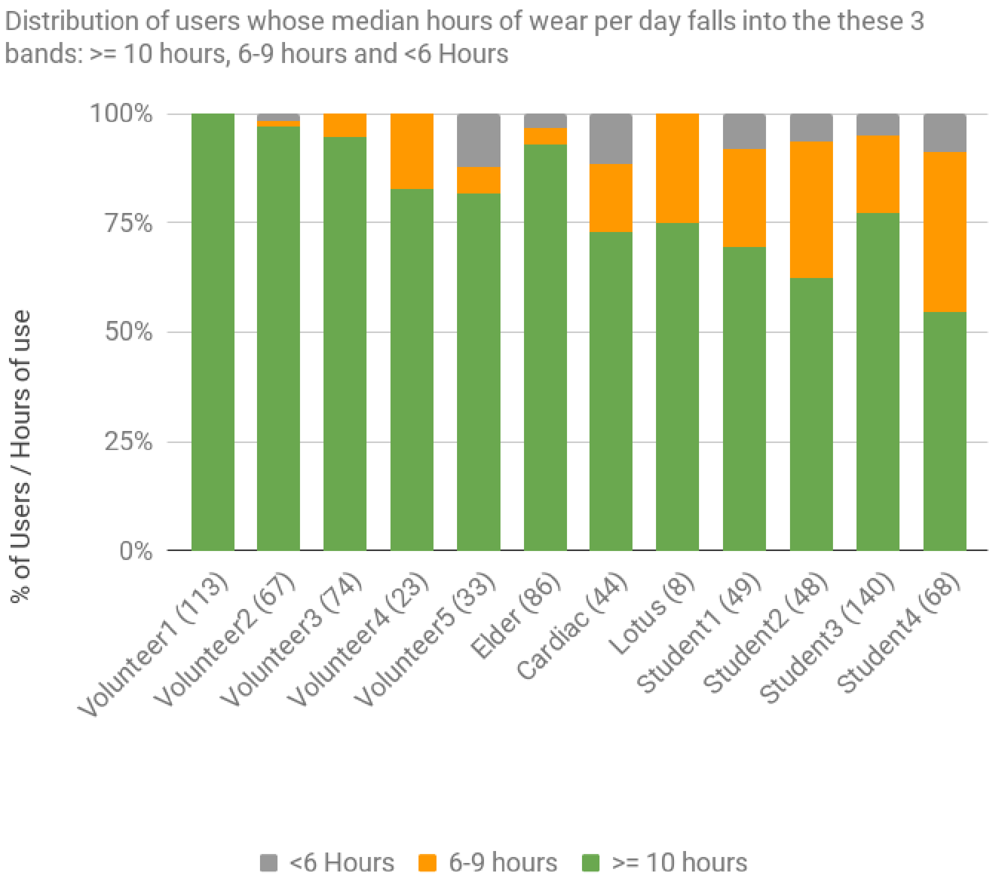


Figure 5.3: Comparison of %-age of users with median wear-time >=10 hours, >=6 and <10 hours and <6 hours. N included in X-axis label.

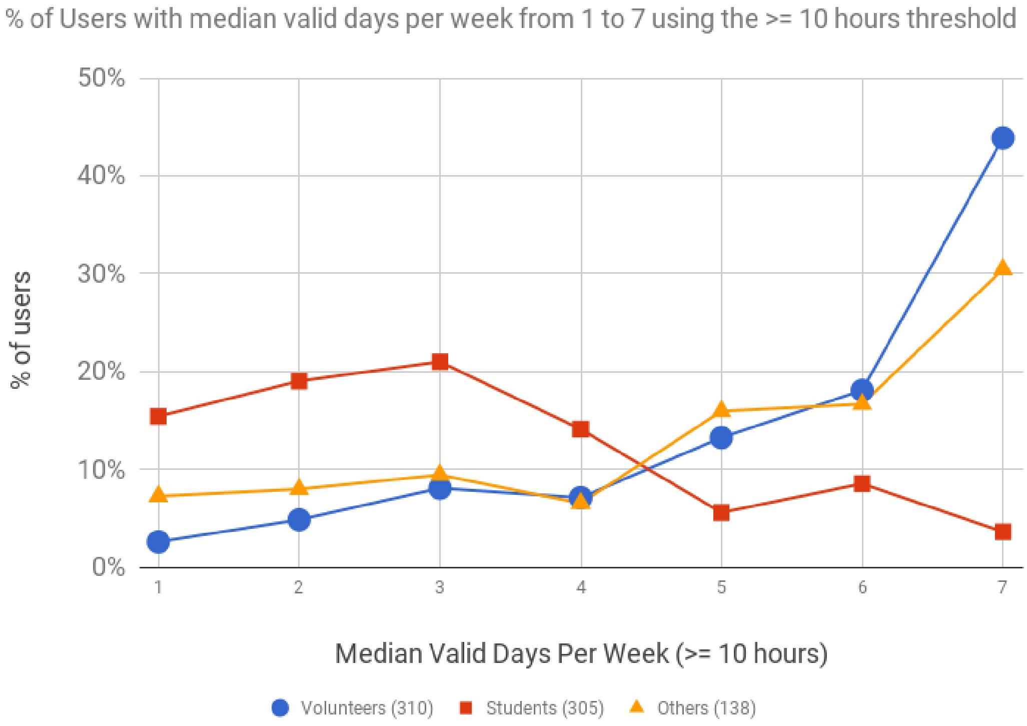


Figure 5.4: Comparison of users median weekly adherence (no. of valid days per week). Grouped as follows: *Volunteers* (*Volunteer1*,*Volunteer2*, *Volunteer3*, *Volunteer4*, *Volunteer5* N=310), Students (*Student1*, *Student2*, *Student3*, *Student4* N=305), Others (*Elder*, *Cardiac*, *Lotus* N=138)

Differences in the People ignored for different adherence measures

We now move from analysis of the *days discarded* and consider the effects for the %-age of *people* whose data is discarded. Figure 5.3 delves into this for the >=10 hour threshold, considering three bands of median wear-time (the count of hours with at least one step). The bottom green part of each bar is the %-age of people with median >=10 hour threshold. The next, orange part enables us to see the potential impact of a less stringent threshold, as it is the %-age of people with median wear-time >=6 and <10 hours and the top, grey part is for < 6 hours. Overall, our analysis shows a very different profiles of wear-time between different datasets. For example, a dataset like the *Student4*, where only about half the participants have wear-time above the threshold of 10 hours, it may be worthwhile exploring less stringent thresholds, and so include more of the population.

Continuing our focus on *people* whose data is discarded, Figure 5.4 shows an analysis of weekly adherence, the number of days a week that were valid (using the 10 hours valid day threshold). The *Volunteers* datasets were dominated by people with high weekly adherence, with almost half of them averaging 7 valid days a week. The Others datasets (*Elder*, *Cardiac* and *Lotus*) have a flatter distribution but still have 30% of people with 7 valid days a week. By contrast, the Students (red squares) have a very different profile of weekly adherence, with very few reaching 7 valid days a week. This is reflected in the weekly adherence across the groups:

- *Volunteers*: mean 5.5 days, SD:1.8, 95% CI: ±0.21.
- Others (*Elder*,*Cardiac*,*Lotus*): mean 4.7 days, SD:2.2, 95% CI: ±0.38.
- Students: mean 2.8 days, SD:1.9, 95% CI: ±0.21.

We repeated this analysis for the least restrictive >0 steps threshold. We found that overall, this more relaxed threshold gave 1.5 (21%) more valid days per week.

5.5.2 Exposing uncertainty: the impact of threshold methods on activity level reporting (RQ2 and 3)

In this section, we show how the adherence measure can affect interpretation of the activity data. We do this in two activity levels, average step counts per day and average active minutes per day.

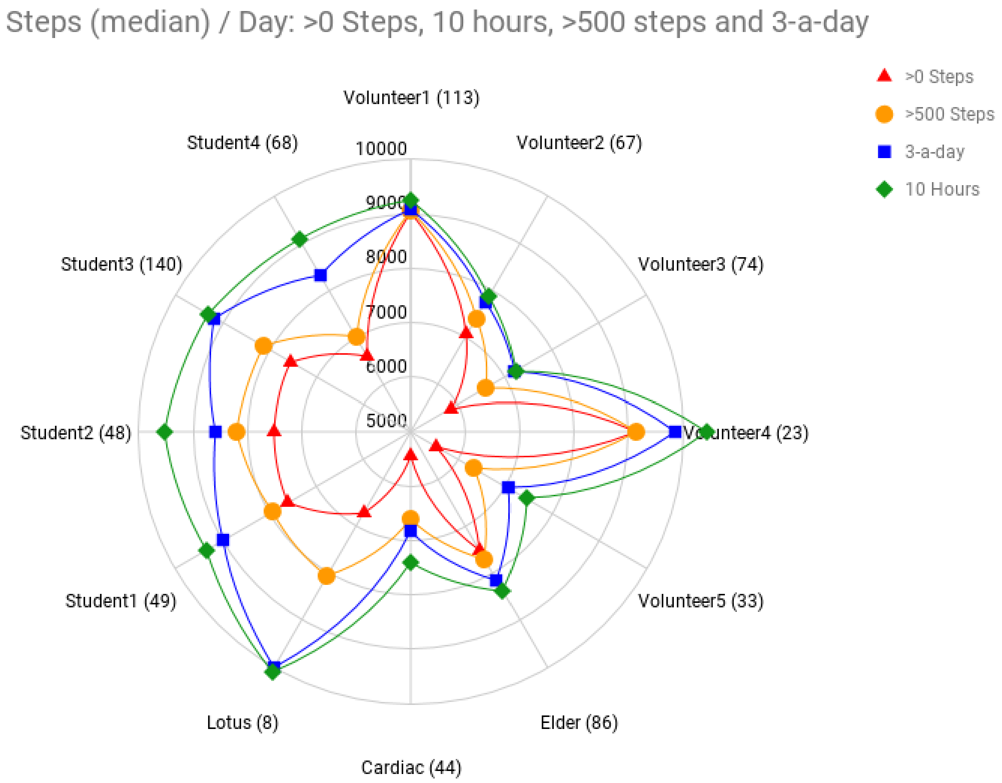


Figure 5.5: Comparison of median steps across populations, showing impact of different valid day thresholds. N included in dataset labels.

Figure 5.5 shows the median steps per day when calculated against the 4 adherence thresholds. For each dataset, it shows, from the top, 10 hours (top end of line), 3-a-day (top of box), >=500 steps (bottom of box) and >0 Steps / day (lower end of line) [1]. This figure indicates how results might differ depending on the adherence definition used. The clear picture that emerges is that the impact varies considerably across the datasets.

For example, across all datasets, the mean steps for each adherence definition are:

- 7133 steps, >0 steps (SD: 1197, 95% CI: ±761)
- 7682 steps, >=500 steps (SD: 926, 95% CI: ±588).
- 8415 steps, 3-a-day (SD: 1060, 95% CI: ±673)
- 8779 steps, >10 hours (SD: 1090, 95% CI: ±692)

There was no significant difference between the two step-count thresholds (>0, >=500), nor between the *through the day* thresholds (3-a-day and >10 hours). However, the step count using the least stringent threshold (>0 steps) is 1,646 less steps (23%) than using the most stringent (>=10 hours). This significant difference is quite large in terms of absolute activity level difference (paired t-test, t_{11}=6.9, p<0.0001, 95% CI: ±527).

We also compared >0 steps and 10 hours thresholds for *Volunteers* datasets and found no significant differences. However, when we examined this with others datasets (i.e., Students plus Others -*Elder*,*Cardiac*,*Lotus*), there was an average of 2,020 steps difference between step counts using the 2 thresholds (paired t-test, t_6=6.9, p<0.001, 95% CI: ±718).

Figure 5.6 shows a similar analysis, now for active minutes per day [2]. Similar to our analysis of steps counts, Students and Others (*Elder*,*Lotus*) datasets had significantly higher active minutes (1.9 minutes more) using the 10 hours threshold compared to the >0 steps threshold (paired t-test, t_5=5.9, p=0.002, 95% CI: ±0.81).

[1] We used a candlestick-like visualisation (or box plot) to convey the spread between the datasets. While it is theoretically possible for steps count for the 3-a-day threshold to be higher than 10 hours or lower than the 500 steps, this is not the case in any of our datasets. So, this format gives a compact summary of our analyses.

[2] To calculate active minutes calculations, we used 120 steps per minute, commonly used to calculate moderate-to-vigorous physical activity (MVPA) [Tudor-Locke et al., 2011]. Some datasets were excluded from this analysis due to lack of per minute activity tracker data (i.e., *Volunteer2*, *Volunteer3*, *Volunteer4*, *Volunteer5* and *Cardiac*)

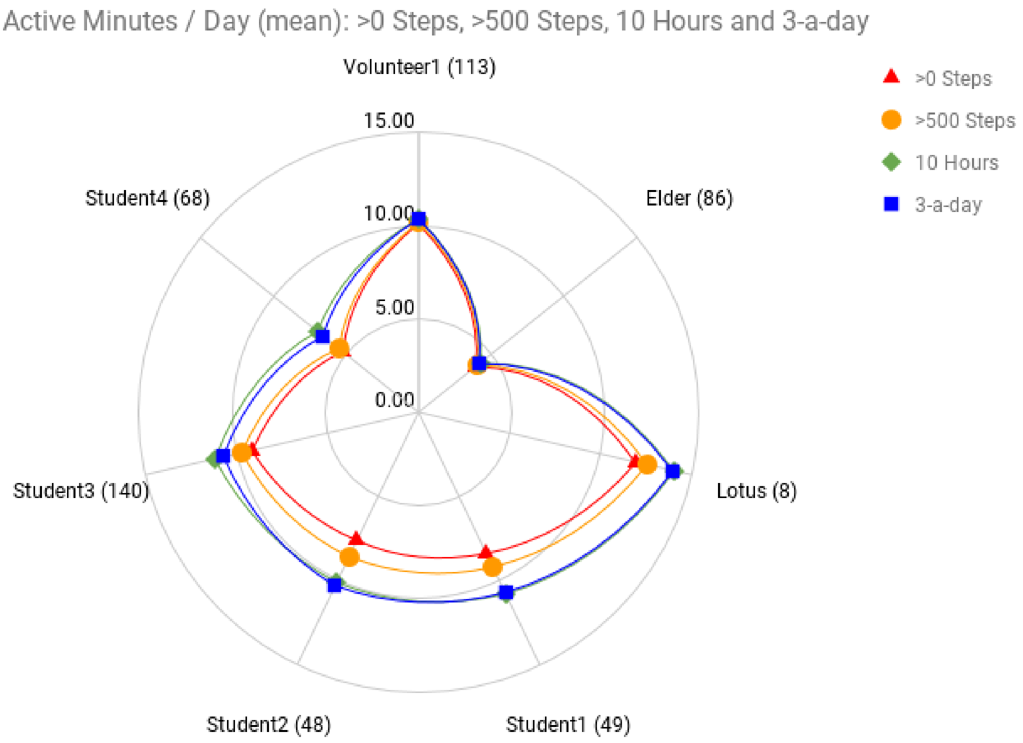


Figure 5.6: Comparison of active minutes results across populations, showing the impact of different valid day thresholds. Datasets *Volunteer2*, *Volunteer3*, *Volunteer4*, *Cardiac* and *Volunteer5* are not included due to lack of per minute data. N included in dataset labels.

5.6 Discussion

In the introduction, we presented examples of core Activity-level and Goal-met questions, both for individuals and in aggregate. As a foundation for our discussion, we now introduce the following questions – these refer to two hypothetical long-term data sets, called *Dataset1*, *Dataset2*.

1. In *Dataset1*, what were people's average daily steps counts in 2014 and 2017?
2. Were people in *Dataset1* more active than those in *Dataset2* in 2017?
3. For the goal of 120 minutes a week of moderate activity, what percentage of *Dataset1* people met the goal in 2014 and 2017?

The first is an Activity-level question, to compare activity level *within* a dataset. The second question is similar to the first, but involves comparisons *between* datasets. The third is similar to the first, but it is for a Goal-met question. These go beyond the analyses we have reported but are useful for broadening the scope of our discussion, building on the reported work. We refer to these questions in the discussion, which starts with the key insights for our three research questions. Building from this, we present a set of recommendations for analysing physical activity data. We then briefly discuss the diverse goals for activity tracking. Finally, we discuss the limitations of our work, along with future directions for research to support systematic and effective accounting for adherence in physical activity data and other personal sensor data.

5.6.1 Key insights for the research questions

Our study showed very diverse impacts of the four core adherence measures across the 12 datasets, with *significant* and *important* differences in *%-age of days discarded* and *%-age of people ignored* (RQ1).

In terms of the number of days of data discarded by the thresholds (Figure 5.1):

- both minimum step-measures, >0 days and > 500 steps were similar for most (but not all) datasets;
- the through-the-day measures, >10 hour and 3-a-day, had diverse impacts – similar on some datasets, quite different for others;
- both through-the-day measures ignored more data than the minimum step-measures.

- these through-the-day measures had quite diverse effects across the datasets.

Some of our datasets had consistently high adherence on all these measures. This was the case for *volunteers* datasets and *Elder* (shown in Figures 5.1 and 5.3). For these populations, strict adherence thresholds cause very little loss of days or people.

Our 12 datasets represent considerable diversity in terms of many factors, including the way the data was acquired, the purposes for which it was collected, the ages of participants, the size of the cohort and the duration. Some trends in the adherence levels are:

- The volunteers datasets tended to have higher adherence than the others. This seems likely to be due to the recruitment methods, tending to attract committed trackers.

- The student datasets tended to have the lowest adherence, perhaps because of the complex factors affecting the students' enthusiasm for tracking and interest in their physical activity levels: they were invited to participate as part of class practical work and the Fitbits were on loan, just for the study period.

Based on the diversity of results across the datasets, and within the groupings (volunteered, students and others), the main observation is that one needs to analyse any new dataset, using all four adherence criteria, to see the impact of each. Our results cannot be used to predict the adherence to be expected for a new dataset.

This impacts calculations of Activity-level questions (RQ2). For example, in the combined student datasets, the >0 steps adherence measure gave a daily step count of 6,952 where the >=10 hours gave 9,423, 35% higher (see Figure 5.5). From a health perspective, this is an important difference. Corresponding to this, the different adherence measures had very different impacts on the data ignored (RQ1). The %-age of *days ignored* moved from 32% to 60% (see Figures 5.1 and 5.5). In terms of people ignored, the >10 hours per day measure excludes more than a third of students (34%), compared with days with >0 steps (see Figure 5.3). For weekly adherence, less than a third of students (32%) averaged 4 days or more per week with >10 hours (see Figure 5.4). This highlights the challenge in interpreting such data to draw conclusions about Activity-level and Goal-met questions. It indicates the potential for introducing *bias* by using an adherence measure that excludes many low adherence users. There is a similar picture for Goal-met questions (RQ3), For example, for the combined students dataset, the >10 hours threshold would ignore 43% of days. our work suggests that there are no easy answers to interpreting such questions but it does point to the importance of reporting adherence measures and their impact along with inferred answers to questions.

5.6.2 Recommendations for systematic analysis of activity data

1: Establish a suitable set of adherence measures to consider

This first step in analysing physical activity data calls for identifying potential adherence definitions. This paper focused on these:

- >0 steps – the least stringent measure;
- >500 steps – to exclude days with very little data;
- >10 hours – a very stringent measure requiring wear for many hours of the day;
- 3-a-day – similar to >10 hours, but requiring wear in the 3 time periods.

We recommend that these four be considered because they range from minimal to quite stringent and these have been used in previous work, providing an initial set of comparative data. There are many other possibilities, and we have touched on some which should also be considered, such as weekly adherence >= 6 days a week, more complex weekly adherence of at least 1 weekend-day and at least 4 weekdays.

2: Explore the impact of the adherences measure(s)

For each adherence measure considered, replicate our analyses to assess its impact in terms of:

- the %-age of *days* ignored
- the %-age of *people* ignored.

Then follows the actual analyses of the data, based on the chosen set of adherence measures, to determine the answers to the core questions, such as the examples at the beginning of this section.

3: Report adherence along with results of data analysis

For many contexts, analysis of tracker data needs to provide a single answer to a question. Even in these cases, we recommend that this result is reported along with:

- the adherence measure used;
- the results of the analyses for this measure in Step 2;

These recommendations provide a way to enhance trust and confidence in information by presenting information about accuracy. In the case of research reports, where more information can be provided, we also recommend providing an explanation for the choice of the adherence measure as well as details of the fuller analyses from Step 2. Over time, this would make it easier for researchers to compare reported results across the literature.

Examples of accounting for adherence in reporting personal informatics results

As discussed earlier, failure to properly account for incompleteness can lead to a loss of trust and ultimate usefulness of such systems [Fogg, 2003, Consolvo et al., 2014, Bentley et al., 2013]. We now consider how our work can help address this problem by enabling a user to consider the impact of adherence when they try to interpret their physical activity data. For this case, we recommend making it possible for the user to:

1. see the adherence measure used;
2. have the opportunity to select the adherence measure the user prefers, based on their own knowledge of their adherence.

Here are two examples of reporting activity levels along with adherence measures.

1. You had 4,500 steps yesterday **in 6 hours with data.**
2. You averaged 9,500 steps a day in the last month **based on the 17 days that you had 10-hours with step data.**

The first example illustrates a way to report steps when a user did not have enough data to qualify as a valid-day (e.g., only 6 of 10 hours). By adding the adherence information, a user can see an indication that the step count may be an under-estimate.

The second example illustrates how to report data over a period of time. The next two examples apply these principles for comparisons.

3. You were less active this summer than last year. **Based on days with at least 10 hours of wear. This is 20% of all your summer days.**
4. You are in the top 20 percentile of your peer group, by age and gender, **Based on days with at least 10 hours of wear. This is 50% of your days and 30% for the peers**.

These examples enable a user to decide whether they trust the result, based on the %-age of days included and their beliefs about their tracker use.

The next example illustrates a goal-met result:

3. You met your goal of at least 30 active minutes a day on 3 days last week. **You were under it on 2 days that had 10-hours of data and the other 2 days had less than 10-hours of data.**
4. You are in the top 20 percentile of your peer group, by age and gender, **Based on days with at least 10 hours of wear. This is 50% of your days and 30% for the peers**.

5.6.3 Reasons for collecting and analysing activity data and implications for accounting for adherence.

Motivations for collecting tracking data

There are many possible reasons for a person to track physical activity, producing a dataset. Tracking may be *initiated by the individual*. Increasingly, smart-phones automatically collect activity data, often without the owner being aware. But much tracker data is purposefully collected. This form of tracking has been the focus of Ubicomp's *Personal Informatics*, as well as the Quantified Self communities [Li et al., 2012, Consolvo et al., 2014, Rooksby et al., 2014, Rapp and Cena, 2016, Choe et al., 2014, Choe et al., 2017, Elsden et al., 2015]. That research reports various motivations for such tracking, such as self-monitoring, self-reflection and conducting n-of-1 experiments [Daskalova et al., 2017, Choe et al., 2017] and with suitable interfaces, to support behaviour change [Consolvo et al., 2014]. (Our volunteer data is most similar to this work.) Beyond this, tracking may be initiated as *part of an* *intervention* (*Elder*, *Cardiac* and *Lotus*) perhaps on advice of a medical professional or in a *study of a population* (like our student datasets). People may use trackers for short periods, for example just for a week to establish a baseline and then at later points, to assess the effects of the intervention, as in [Tiedemann et al., 2015]. Others track consistently over long period. In all these cases, the answers to questions such as the examples at the beginning of this section demand a solid adherence analysis such as we have recommended. This can be the basis for assessing the accuracy of the results and if multiple adherence measures are used and reported on, this will support comparisons in the literature.

Using adherence measures to change wearing behaviour versus using adherence measures to make sense of available data

If adherence measures are reported along with results, as in the examples above, this may give individuals information to help them re-consider their adherence levels. For example, Tang et al [Tang and Kay, 2017] reported that some of their participants said that they planned to be more adherent. But a core goal of our work is to provide foundations to harness available activity data, regardless of whether the adherence is low, intermittently high, consistently high or any other pattern. Even with lower adherence, if the adherence is reported, people may well gain valuable answers to their questions, along with information to assess the reliability.

5.6.4 Limitations and future work

The driver for our study design was to provide foundations for accounting for adherence and this drove decisions about each element of the study design. We now discuss the limitations of our work. If our recommendations are followed, we hope to see future work adding to our results for each of these aspects and we give pointers to these. We conclude with future directions for smart-phone data, drawing on our work.

Questions explored

Our three research questions explored the impact of adherence measures and how to account for adherence when answering questions about physical activity. We began with questions about pure Activity-level and Goal-met questions. At the beginning of this section, we introduced similar questions for comparisons *within* and *between* datasets. These are basic questions, representing just a starting point in analysing physical activity data.

Adherence definitions and analyses explored

To manage the complexity of results to present, we carefully selected just 4 valid day definitions, two each, *minimum-step-count* and *through-the-day*. In the case of minimum-hours-in-the-day, we focused on 10-hours a day because it is common in literature, but we did explore the impact of considering <6 hours and 6-9 hours. Similarly, the set of experiments we report was carefully chosen to explore our research questions and enable a reader to see the key results showing the importance of definitions of adherence.

Datasets: participants

while our datasets are large and diverse, they all do come from authors of the paper. It will be valuable to see this work replicated on other data sets. Our results show that our volunteer and student categorisation does show some commonalities within these categories. There are many dimensions that may be important for describing key characteristics of the people in a dataset, such as gender, age, health status and importantly motivation for collecting the data. Associated with these are characteristics such as the duration of data collection.

Datasets: activity tracking devices

Our datasets all come from similar trackers, although people had different variants on these trackers. None of our devices can distinguish inactivity from non-wear. Many current tracker devices can do this (for example, [Jeong et al., 2017]). Even in our datasets, some participants may have changed devices through the time. We ignored these issues in our analyses. To answer the three questions at the beginning of this discussion, device differences need to be considered. There is also a broader discussion around accuracy of trackers for reliability and comparability [Migueles et al., 2017, Evenson et al., 2015, Trost et al., 2005, Tudor-locke, 2016].

Using Smart-phone physical activity data

Another important *wearable* (or carried) device is the mobile phone. Our work provides a foundation for identifying meaningful measures of adherence for tracking physical activity as captured by a mobile phone. For people who always wear/carry their phones when awake, and so having high adherence, the phone could provide a reliable way to track activity. However, many people do not do this. Yet a recent large study assumed it was reliable to compare populations, based on assessing daily wear-time from the first to last phone use in the day [Althoff et al., 2017]. On this basis, they reported a mean step of 5,039 across 111 countries and 717,527 users and compared step counts by country. There is no indication of incompleteness in their data. Are comparisons between different countries based on comparable daily adherence (e.g., 80% US versus 70% Australia or 80% US versus 20% China)? Our work suggests that it would be valuable to also calculate the step counts for other adherence measures that are well suited to a mobile phone. An adaptation of our $>$10 hours measure seems a good starting point.

Our work has implications for other inferences from wearable devices that measure many other things, such as heart-rate, stress, sleep quality, air-quality. We see an important role for adherence measures in the ubicomp field when applications seek to combine and provide self monitoring using data from different classes of devices and over time. As tracking capabilities, devices and sources of data change, so too would the wearing and adherence patterns of the data people have collected. This is true for both aggregate data and personal informatics.

5.7 Conclusion

We reviewed both health and computing literature to establish definitions of *adherence*, a measure of incompleteness, to help answer two important classes of health questions: Activity-level and Goal-met.

We also introduced a new adherence measure *Valid-Goal-Day* needed when answering Goal-met based health questions. Our analysis of 12 large and diverse physical activity tracker datasets showed that previous threshold based methods of addressing incompleteness is not appropriate for large scale *volunteers* datasets such as those collected through personal use e.g., Fitbit users. When dealing with *volunteers* datasets, we recommend to analyse and also to report adherence measures for individual applications. Further research is needed on how the information can best be delivered. For aggregate reports, it is also important to report adherence criteria used along with the adherence measures and their impact on the activity level results.

Chapter 6

Implications for personal informatics applications

Preamble

This work is currently pending submission. It reports on contributions described in section 1.2.4.

6.1 Understanding physical activity tracking data: wear-time matters

Physical activity tracker is widely used. However, as people do not wear their tracker all day, every day, their long-term data is typically incomplete. We show how key interfaces fail to account for that incompleteness and why this is a problem. We present several potential solutions.

Physical activity is critical for good health. Yet many people currently fail to get enough to meet health recommendation [Haskell et al., 2007]. Physical activity trackers have the potential to help change this. Part of the reason for this is that they are increasingly available. Millions of people have dedicated activity trackers. For example, in the US alone, 18% of adults own a wearable fitness tracker [Albert, 2017]. Even more people have smart-phones or smart-watches that can track physical activity. This means that large proportions of the population can readily collect fine-grained data about their physical activity as they go about their normal lives. This trend is likely to continue, both in the rich and developing world.

However, activity tracker data can only be useful if interfaces present it in a manner that enables them to gain meaningful insights. To do this, it should enable people to answer questions such as:

- *Have I met recommended levels of physical activity in the last week?*
- *How much physical activity did I average over the last week? Or last month? Or year?*

While knowledge alone is not enough for substantial behaviour change, it is a critical foundation. If the answer to the first question is no, a person could become aware that they should change their behaviour. If the answer to the second question is 3,000 steps a day, this provides the basis for setting a target that is achievable, a core for successful behaviour change, based on S.M.A.R.T goals [Locke and Latham, 2002a]. There are also other questions that activity data should enable people to answer, such as:

- *How am I doing compared to other people like me?*

- *Did my move from the city to the suburbs a year ago impact my physical activity levels?*

The first is a comparative measure that may be helpful when a person is reviewing their target goal. The second illustrates a case where a person's long-term data should enable them to gain insights about the impact of major life changes on their physical activity. We now show how widely available current interfaces for activity data provide answers to such questions. We highlight a serious deficiency in these interfaces. This is because they present information that may be misleading because the users ignore wear-time for the device. This is important because many people do not wear their trackers or carry their phones all day, every day. Since interfaces ignore this, they fail to account for the inaccuracy due to incompleteness of available data when people have not worn their tracker or carry their phone.

6.2 The Dangers of Incomplete Data

When data is incomplete and this is not accounted for properly, people may not trust it and ultimately this compromises its usefulness. Fogg [Fogg, 2003] warns that when systems produce questionable data, people are less likely to trust them and that this makes them less useful as a behaviour change tool. Consolvo et al ([Consolvo et al., 2014], page 234) also recognised the importance of missing data for providing self-monitoring feedback. They reported how presentations that fail to account for missing data can incorrectly suggest trends and patterns. The Health Mashups system [Bentley et al., 2013] was designed to integrate data from multiple sources to drive behaviour change. The authors highlighted problems when data was sparse or incomplete because some users noticed contradicting information and inaccurate recommendations. They note that this led to a loss of trust in the tool.

In the case of physical activity tracker data, incomplete data sets are common. This is because many people do not wear their trackers all day every day. Wear-time is important since the trackers only capture steps when worn. For phones, the problem is different but even more serious since many people do not carry their phones all the time. A growing body of work has studied wearing patterns for dedicated trackers [Tang et al., 2018, Tang and Kay, 2017, Tang and Kay, 2016, Meyer et al., 2017, Epstein et al., 2016b] , studying how wearing patterns vary significantly between people and over time and that there can be large gaps in long-term activity tracker data. To understand this better, we analysed 12 diverse datasets from 753 users, with over 77,000 days of data from wearable tracker devices [7]. We found that these people had many data gaps - a total of 73,000 days without data - a similar number of gap days to data days. Importantly, we found that different ways of dealing with the incompleteness of the data had a significant impact on the answers to questions like those we posed earlier, about how active people were and whether target activity levels. The core conclusion was that they it is critical to take account of wear-time when interpreting physical activity data. We now show that some of the most widely available physical activity interfaces fail to do this.

6.3 Hiding in Plain Sight: wear-time

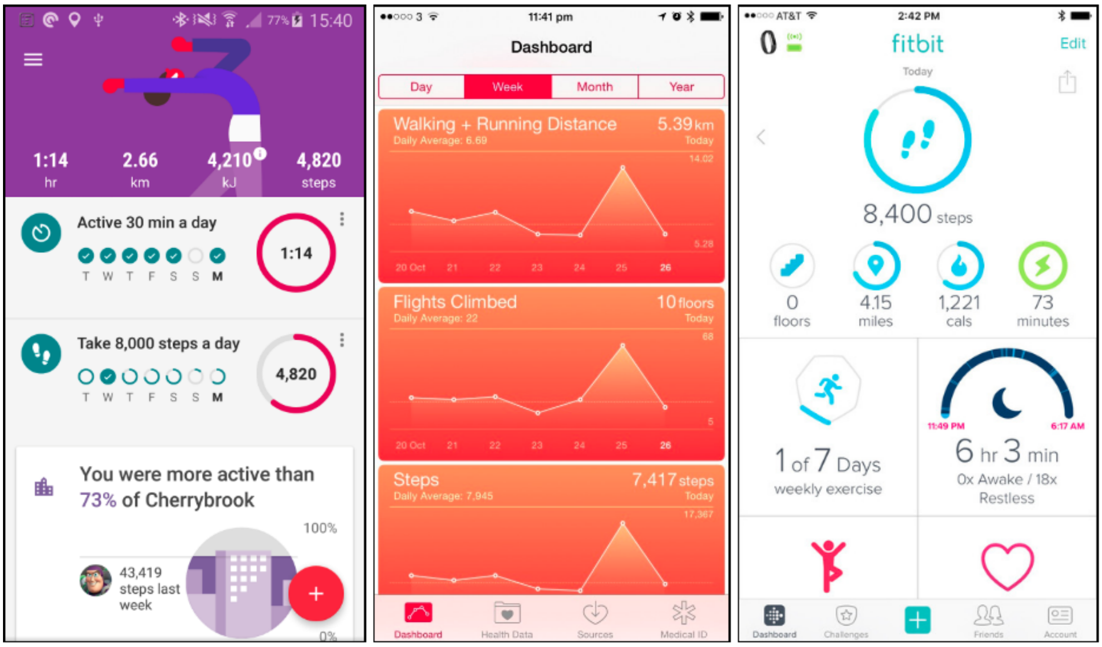


Figure 6.1: Examples of physical activity dashboards for key platforms: Google Fit app (left), Apple Health (middle) and Fitbit (right). Note: screenshots are not all the same user.

While it seems obvious that calculations of tracker data should take account of incompleteness of the data, widespread applications fail to do this. Figure 6.1 shows examples of dashboards from three key activity tracking platforms. We selected the two largest smart phone platforms: Android with Google Fit; and Apple OS with Apple Health. We also taken the FitBit as it has an important place as a supplier of dedicated activity trackers, both in terms of early entry into the market over 7 years ago, and their broad uptake.

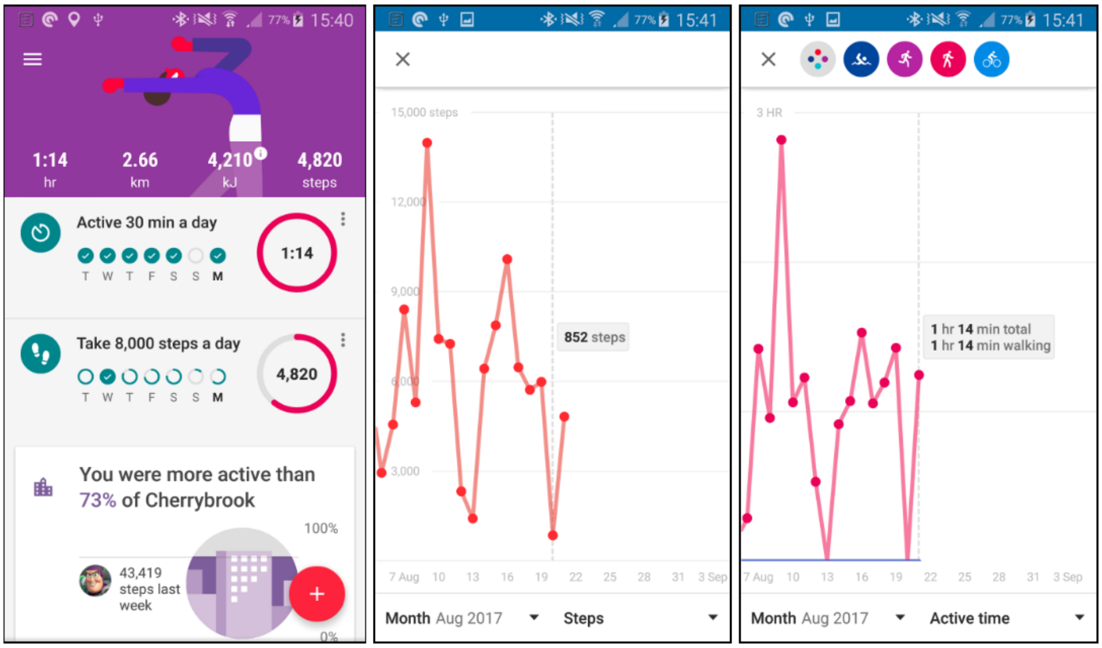
Figure 6.1 (left) shows the Google Fit interface for a user tracking two goals: 30 active minutes per day and 8,000 steps per day. This shows that the user met their 30 active minutes goal on all but 1 of the last 7 days. However, the user failed to reach their 8,000 steps target on 6 of the last 7 days. Figure 6.1 (middle) is an Apple Health App weekly chart showing the daily step counts over the last week as well as stairs and distance travelled. Figure 6.1 (right) shows the Fitbit dashboard which, in addition to steps and active minutes, has exercise done, (1 of 7 days), sleep and calories information. Notably, none of the above interfaces gives any indication of wear-time. None even attempts to acknowledge any uncertainty in the information presented. We will now drill down on the results as they appear in Figure 6.1 (left), particularly that it seems to show that this user:

1. Averaged 4,820 per day
2. Met their 30 active-minutes target on all but Sunday

In Figure 6.2, the leftmost screen is the same as in Figure 6.1 but now, alongside it, the middle screen shows an additional screen available to show the daily steps for recent days and the right one shows the active minutes per day. This additional data might help a person remember how their recent wear-time affected the results. For Sunday 20 August, with step count just 852, the user could consider:

- *Was this low because I was very inactive?*
- *or was it low because I did not wear their device much that day?*

If this was the last week, the user may recall their wear-time. So they may recall that they actually wore the track only briefly on Sunday and that for the rest of the week, they wore the device only for a morning walks, achieving the 30 active minutes a day. Then they could conclude that the 4,820 step per day is an underestimate of their overall activity that week.



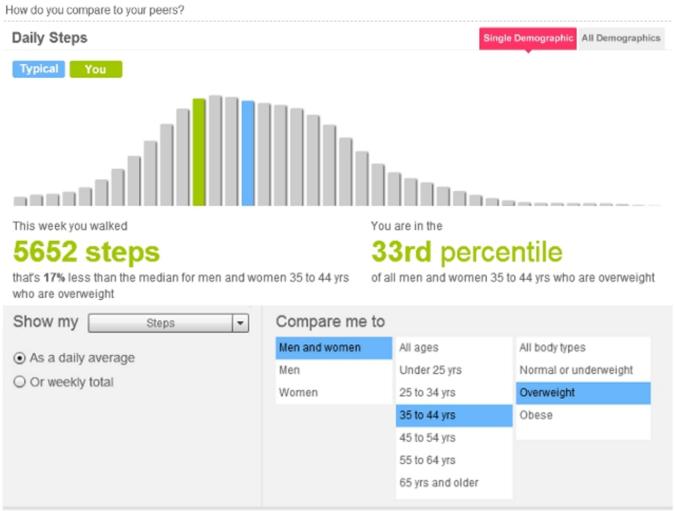
You were more active than 73% of Cherrybrook

Figure 6.2: (left) dashboard summary as in Figure 6.1, (middle) detailed steps/day, (right) detailed active-minutes/day.

A person may remember this for the last week. However, if they look at older data, to understand their longer-term activity levels and answer the questions above, this is a problem. They many not recall their wearing behaviour a year ago - or even a month ago. Once they forget their wearing behaviour, they simply cannot know how much to trust the step count data as displayed here.

We now consider the lower part of the Google Fit dashboard which reports this user was more active than 73% of the suburb they live in, Cherrybrook (at the left in Figures 6.1 and 6.2). The user may consider this is a useful comparison for them. But this, too, ignores wear-time for both this user and the Cherrybrook population. While the user may well remember their own recent wear-time, they have no way to determine the wear-time of the Cherrybrook population. So, the user cannot judge the reliability of this comparison.

In Figure 6.3, we show another example of comparative data, this time from a premium Fitbit service. This allows a user to refine the comparisons they want to see. In the figure, it is based on comparing daily steps, against the population of 35-44 year olds who are overweight. This user is reported as being at the 33rd percentile -



This week you walked **5652 steps** that's **17%** less than the median for men and women 35 to 44 yrs who are overweight

You are in the **33rd percentile** of all men and women 35 to 44 yrs who are overweight

● As a daily average

○ Or weekly total

Figure 6.3: Example screenshots of a comparative report can be available from Fitbit.

well below the median user in this group. Here again, as in the examples above, the user has no way to know about the wear-time of the peer group because the interface fails to make wear-time available. So the user cannot assess the meaningfulness of this comparison.

In summary, this section introduced examples of key commercial interfaces for physical activity data. None of these enable a user to see the impact of wear-time on results presented. Studies of wear-time indicate that most people have incomplete data [Tang et al., 2018, Meyer et al., 2017, Epstein et al., 2016b]. This means that such interfaces make it very difficult for a user to judge their long-term physical activity level and to compare it with health recommendations, with a peer-group and how their activity at one time compares with that at another time.

6.4 What can we do? How interfaces might account for wear-time.

In this section, we present ways to overcome the problems just described. We illustrate three strategies to make wear-time information available along with activity data results. For simplicity, we illustrate this for a person who considers step counts reliable only when they had at least 10 hours of wear-time.

6.4.1 Strategy 1: Interfaces can report calculated step counts along with wear-time.

Consider Figure 6.1 (left), which shows last weeks' step count as 43,419 (at the bottom). If a system simply divides this by 7 this would be reported as 6,203 per day. But this includes days with very low wear-time, like the 852 steps in Figure 6.2 center. Taking account of wear-time, the interface could report:

> *You averaged 7,095 steps last week (for 6 days with >10 hours wear-time.)*

This does two things. It gives a more reliable estimate of step count. It also indicates the wear-time threshold. This strategy can be used for long-term reports, such as:

> *You were less active this winter than last year*
>
> *(for days with >10 hours wear-time 30% of days last winter and 50% this winter)*

In this case, the user may not trust the comparison once they can see the low proportion of days with 10-hours wear-time and the differences between these. Alternatively, they may just interpret this result more cautiously because of this information. Similarly, when comparing with peer groups, as in Figure 6.1 (left) the wear-time could be reported as in:

> *You are in the top 73% of Cherrybrook*
>
> *(for days with >10 hours wear-time - 36% of days for you and 65% average for Cherrybrook)*

Now the user can see that their wear-time time is below the peer group and that the simple comparison might not be accurate. Another way to report the impact of wear-time is:

> (*Your average wear-time was 6 hours a day, below the Cherrybrook average of 12 hours).*

This also enables the user to consider the impact of wear-time.

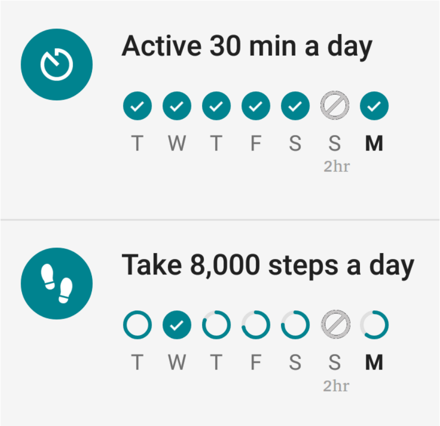


Figure 6.4: Distinguishes days where wear-time is below 10 hours.

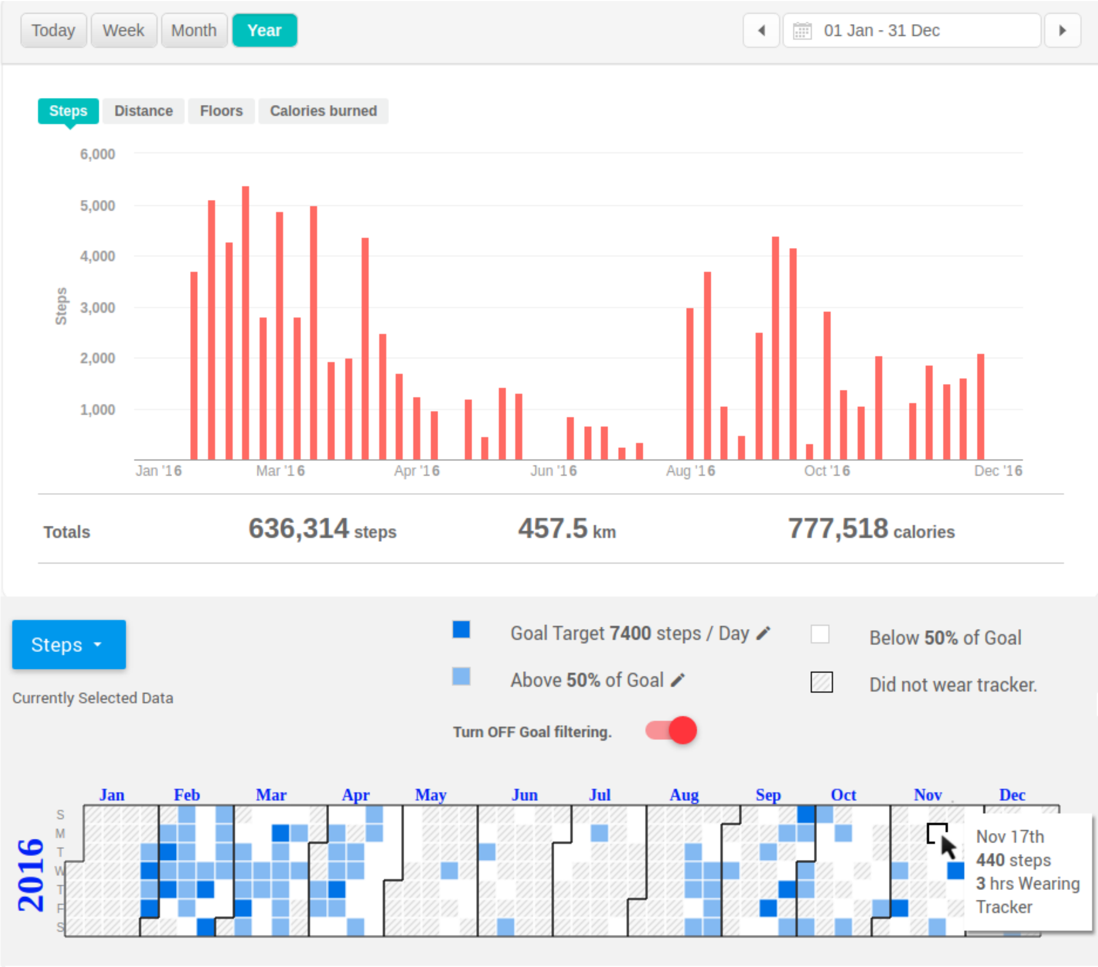
6.4.2 Strategy 2: Interfaces can report low wear-time if a daily target is not met

Figure 6.4 illustrates one way to report wear-time information along with activity levels. The original visualization, in Figure 6.1, has been adjusted grey out any day with low wear-time, reporting the actual amount of wear-time. In the top part, for 30 active minutes, the missed target on Sunday is now grey and shows just 2 hours of wear-time. The lower step count display shows the target was only met on one day, Wednesday. It also shows there is not enough data to assess this on Sunday and that the other days really were below the 8,000 steps target.

6.4.3 Strategy 3: Interfaces for long-term activity data can enable users to see the impact of wear-time

The examples above are for short term data, when the user may be able to remember some aspects of their wear-time and they may be able to take this into account when trying to interpret the data for a single day or the last week. For longer term data, people are unlikely to remember details of their wear-time.

Figure 6.5 top shows the current Fitbit.com dashboard. Like the earlier examples of interfaces for the very common commercial tracker platforms, this one takes no account of wear-time. By contrast, the lower interface shows a form of the iStuckWithIt interface [Tang and Kay, 2017]. This shows days where the user met their target in



Totals 636,314 steps 457.5 km 777,518 calories

Goal Target 7400 steps / Day

Above 50% of Goal

Below 50% of Goal

Did not wear tracker.

Turn OFF Goal filtering.

Nov 17th
440 steps
3 hrs Wearing Tracker

Figure 6.5: (**Top**) yearly bar chart screenshot from Fitbit.com. (**Bottom**) iStuckWithIt calendar chart from [Tang and Kay, 2017] distinguishes days with no wear-time.

dark blue, with lower counts in light blue if they are at least half that target and the white which is not. When there is less than 10 hours of tracker data, the days are grey. In conjunction with an interface like this, the user could be advised about changes in their activity level, with an indication of the impact of days they had no data: You were less active in the last quarter of 2017 than the first. (This is based on days with >10 hours data – 70% of days in the first quarter, 40% of the days in the last quarter. The iStuckWithIt interface was designed to enable users to readily distinguished days they did not wear their tracker. In the evaluation user study, participants were readily able to interpret this aspect. The interface also provides a details-on-demand mouseover so that the user can see the actual wear-time on any day (e.g., cell of date 17-Nov-2017, bottom right of Figure 6.5).

6.5 Conclusion and call to action

Physical activity is critical for good health. The huge uptake of dedicated trackers reflects the potential that people see for activity trackers to help them achieve healthy levels of activity. Similarly, the widespread availability of phone-based activity tracking has the potential to make this data available to huge populations. But wear-time for phones needs to be considered when reporting activity.

Already, some FitBit users have years of data; increasingly, many people will have long-term activity data that should be valuable because it could give them insights about their long-term activity: how active they are, whether they meet public health recommendations, how their activity levels have changed over time and how they compare to relevant peer groups.

This paper is a call to action for better reporting of data from worn or carried physical activity sensors. For meaningful presentation of activity data, it is important to take account of wear-time. The tracking devices have the timing of data to calculate wear-time. That information needs to be available to users so they can understand their activity data, so that they can make meaningful interpretations of their data, taking account of wear-time. We have recommended three strategies. The simplest is to report wear-time along with step counts and activity level in interfaces. The second is to take account of wear-time when reporting data. For example, an "average" step count should reflect the steps recorded on days with enough data for the user to trust the result. The third approach is to incorporate richer information about wear-time in interfaces to long-term data. There is much work to be done, exploring how to design interfaces that people can readily control and understand. Only then, will we realise the potential of physical activity tracker data be give trustworthy answers to important questions.

Chapter 7

Conclusion and Future Directions

7.1 Contributions

This thesis explored how to make sense of long-term physical activity data, both at the individual and the aggregate or population level. Our work exposed the important and previously under explored challenges associated with data completeness. We then explored the design of user interfaces for self-monitoring and reflection, and finally, we offered adherence measures as a framework for addressing incompleteness challenges when making sense of long-term physical activity data.

Together, our three studies made five contributions, as summarized in the map of the thesis shown in Figure 7.1. [1]

First, in our study of 237 university students' tracker wearing behaviour over a semester, we exposed the significance and potential impact of data incompleteness when reporting on long-term physical activity data.

Our second study of 21 existing long-term trackers made two key contributions. First, this study is the first to examine how they make use of and understand their own long-term physical activity data. Second, our lab study of *iStuckWithIt* showed that our interface design was useful in helping users monitor, and reflect on, their own long-term data. These insights can inform future applications and user interface designs.

Our final study, with the pooling of 12 diverse datasets, demonstrated the limitations of existing methods for dealing with incompleteness in long-term physical activity data and showed that a new approach is needed if we are to make sense of such

[1] Note: Figure 7.1 also appears as Figure 1.1.

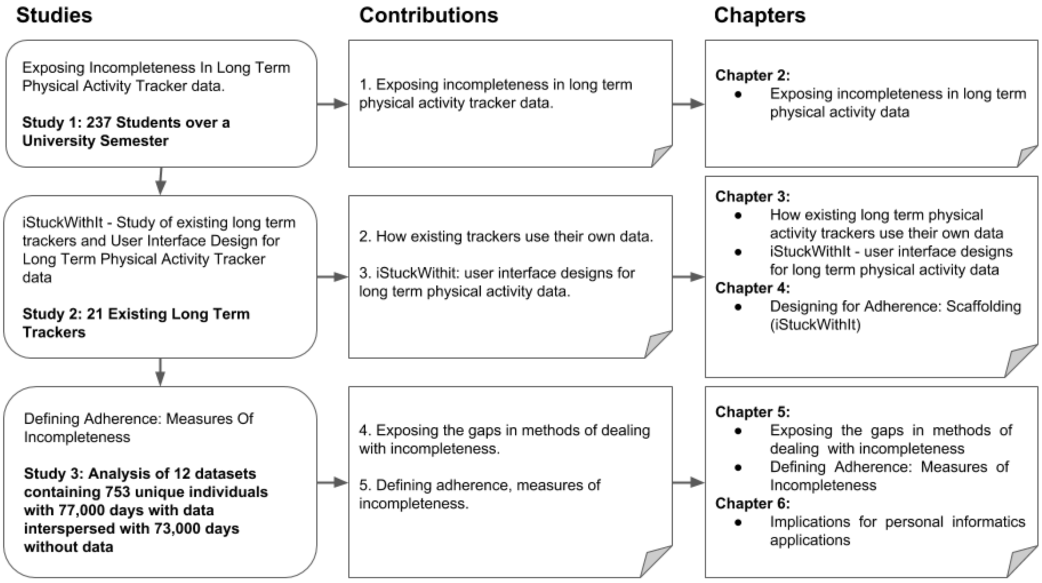

Studies
Contributions
Chapters
Exposing Incompleteness In Long Term Physical Activity Tracker data.
Study 1: 237 Students over a University Semester
1. Exposing incompleteness in long term physical activity tracker data.
Chapter 2:
Exposing incompleteness in long term physical activity data
iStuckWithIt - Study of existing long term trackers and User Interface Design for Long Term Physical Activity Tracker data
Study 2: 21 Existing Long Term Trackers
2. How existing trackers use their own data.
3. iStuckWithit: user interface designs for long term physical activity data.
Chapter 3:
How existing long term physical activity trackers use their own data
iStuckWithIt - user interface designs for long term physical activity data
Chapter 4:
Designing for Adherence: Scaffolding (iStuckWithIt)
Defining Adherence: Measures Of Incompleteness
Study 3: Analysis of 12 datasets containing 753 unique individuals with 77,000 days with data interspersed with 73,000 days without data
4. Exposing the gaps in methods of dealing with incompleteness.
5. Defining adherence, measures of incompleteness.
Chapter 5:
Exposing the gaps in methods of dealing with incompleteness
Defining Adherence: Measures of Incompleteness
Chapter 6:
Implications for personal informatics applications

Figure 7.1: Map of thesis: Studies conducted, the contributions of these studies and the chapters that report on these studies and contributions. Note: this is a copy of Figure 1.1, repeated here for convenience.

data.

Our final contribution was our definition of adherence; a way to measure completeness in physical activity data. We offer adherence measures as a framework for dealing with completeness in tracker data and have provided guidelines with examples of how these measures could be used.

It is important to highlight that the final study took a step back and systematically considered how to define adherence measures as a framework; it was the culmination of the findings gained from dealing with incompleteness in long-term physical activity data across all three studies. Moreover, all three studies contributed towards demonstrating the importance of adherence measures when making sense of long-term physical activity data. For example, in Study 1, daily and hourly adherence measures helped us to expose significant differences in wearing behaviour between student groups. In Study 2, using daily, hourly and weekly adherence measures helped us to highlight significant variations, even when considering data from individuals who have been using trackers for extended periods. In addition, when adherence patterns were exposed to users using *iStuckWithIt*, this information helped the users to reflect on gaps in tracking, and in some cases, to derive meaning from the missing data (e.g., "*I was on holiday this month but was actually quite active*").

7.2 Future Directions

Consumer adoption of trackers, either dedicated or non-dedicated (e.g., smartphone, smart wearables), is expected to grow. It is likely that many, if not most, people will have collected vast stores of personal as well as health-related data that spans years and even decades.

However, while this collection of data will continue to grow in the coming years, there are significant barriers to users making use of this data for their long-term health and well-being goals.

In this section, we look to the future of years and even decades and highlight key questions that long-term physical activity data might be able to answer. We then discuss key barriers that must be addressed before long-term physical activity data can become useful. A key challenge is the need to address incompleteness in long-term data and its centrality to making such data useful [2].

7.2.1 What can long-term physical activity data tell us?

We identify three important classes of questions and illustrate each with examples in relation to physical activity tracking. In the final section, we discuss how this relates to other sensor data types.

Learning about trends and patterns

This is the simplest class of question – it enables a person to discover trends over the long-term, answering questions like these:

- Have I become less active now than I was 10 years ago?
- Am I generally more active on weekends?
- The newest advice is to get 60 active minutes a day - have I been achieving that?

[2] This work was presented as part of "*A Short Workshop on Next Steps Towards Long-term Self Tracking*" at the Computer Human Interactions (CHI) conference in 2018.

Personal Hypothesis

Long-term data also enables people to test their beliefs about factors affecting their level of activity. They may have also tried various strategies to achieve long-term goals.

- I believe that when I moved to the suburbs a year ago, I became more active.
- Taking public transport to work (as I do on Mondays) means I am more active than the three years I had a gym membership.

Remembering and reminiscence

We also see potential and benefits from reminiscence around personal long-term health data. This was shown in a study of 15 long-term users [Elsden et al., 2015] which reported the ways that participants interacted with their data via storytelling and reminiscence. In this case, the question is: *what memories does my data evoke?*

7.2.2 Human-computer interaction challenges and research directions

If we are to be able to answer questions like those above, there are several challenges within the field of human-computer interaction (HCI) that must first be addressed. We first describe the technical challenges and their associated interface challenges. We then discuss core interface challenges for exploring and understanding long-term data.

7.2.3 Combining data from different stores

In the long-term, activity tracker data may be stored by various vendors of tracking technology. This can make it very difficult for users to aggregate all their data and poses a serious problem as devices, companies and technologies evolve. It is not even clear if users own their own data, and in practice, users have very little control over it. For example, users may not know what data is stored (steps/minute versus steps/day) and granularity and detail available is at the vendor's discretion.

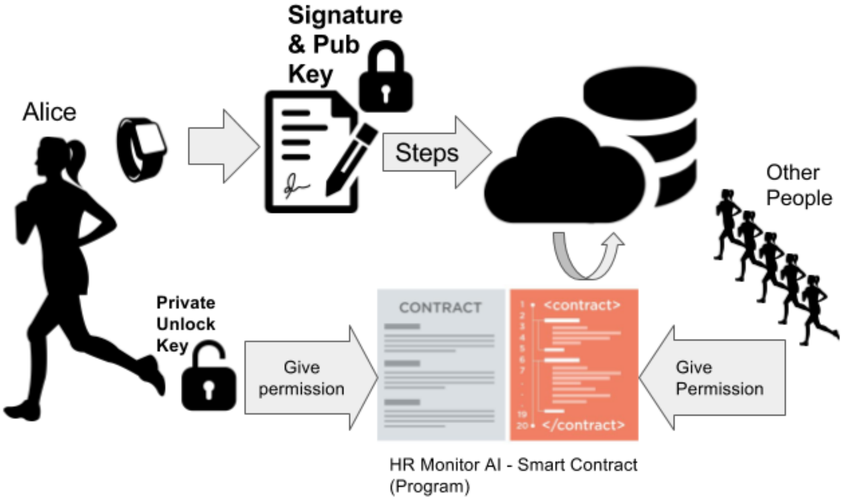


Figure 7.2: Illustrative example of a Blockchain

Data privacy, security, provenance and management

Other major barriers relate to security and privacy [Kay and Kummerfeld, 2012, Althoff, 2017]. Centralised data is easier to analyse but is more vulnerable. To make sense of data, it needs to be associated with the source, in terms of who or what generated the data. If the data is shared, users should also be able to choose who can access it, for what purpose, when and for how long. Kuo et al. [Kuo et al., 2017] proposed the use of blockchain technologies as a promising approach to decentralising trust, managing data provenance and granular control of what and how personal data is used (e.g., smart contracts).

We illustrate this using a hypothetical blockchain scenario that is based on a health record ecosystem example, as shown in Figure 7.2. From 2020, Alice's new smartwatch signs all her physical activity sensor data and her heart rate data with her blockchain signature (a type of key that identifies it is her data and only she can unlock it). It then uploads this encrypted and signed data to a central cloud storage. Alice, along with millions of other people, has given permission for a health AI (developed by researchers from a leading public health institute) to monitor her heart rate, and both contribute to a public health study and provide her with personalised feedback. The monitoring and analysis performed by this AI is executed as a smart contract (a program that is executed on the blockchain to perform a specific function). The health AI creates and sends the report to Alice and to her doctor who she has given access to the report but not the sensor data. There are many more uses and examples of this type of applications. Kuo et al. [Kuo et al., 2017] provide a more detailed review of potential uses and benefits of a blockchain-based health record management system.

- Decentralised management (patient-managed health care records)
- Immutable audit trail (unalterable patient records)
- Data provenance (source verified medical records)
- Robustness/availability
- Security/privacy

It is important to note that while blockchain technology is promising, it is not proven and largely theoretical at this stage. However, it is a good example to use as its supposed benefits and features covers the core barriers that must be overcome for us to fully realise the potential of decades of personal data.

User interfaces to control and manage data

Regardless of the storage technology used, the ultimate goal of any health data management system is to benefit its users. We need to prioritise HCI interface designs to support these potentially useful technologies. For instance, blockchain technology, just like any other complex technology, is difficult to understand let alone trust as the keeper and guardian of our potentially highly personal and sensitive information. For people to trust it and make use of it, there need to be interfaces based on effective design insights that overcome barriers from a user's perspective.

It is clear that the technical challenges of storage and privacy management have parallel interface challenges. Studies of online social network users point to how difficult these challenges are, indicating the mismatch between intention and actual settings. For example, one study revealed that privacy settings matched users' expectations only 37% of the time [Liu et al., 2011].

7.2.4 Interfaces for interpreting data

Another key challenge is to make the data available in a suitable form. For example, we created a calendar visualisation [Tang and Kay, 2017] like that shown in Figure 7.3.

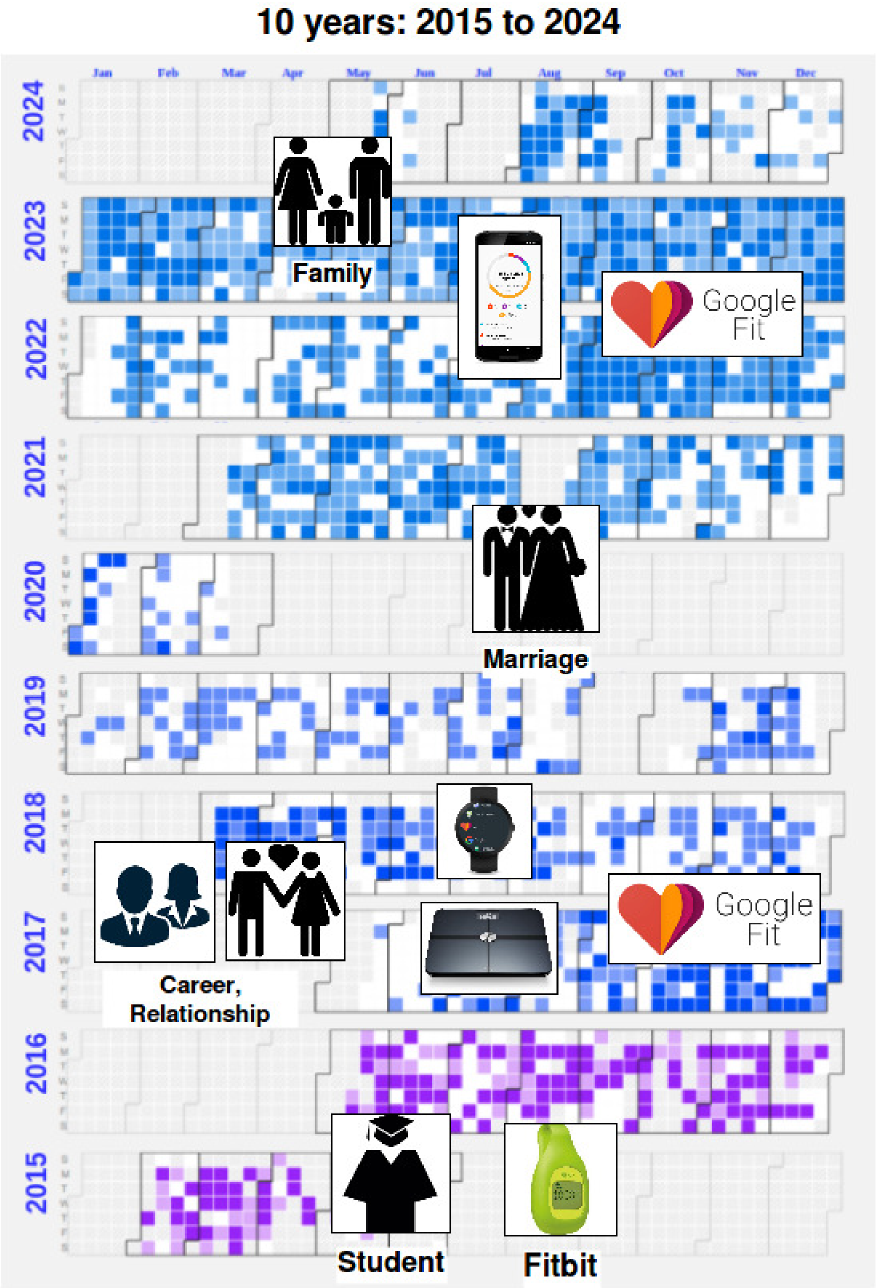


Figure 7.3: Illustrative overview of physical activity data (e.g., steps) from a hypothetical user, Alice, over 10 years, starting in 2015. Calendar chart view (steps/day): dark blue indicates reaching her goal (e.g., 10K steps), light blue indicates 50% of goal, white indicates less than 50%; grey indicates no data. Icons (overlaid) to indicate major life stages (e.g., student, marriage) and devices and vendor used (e.g., Fitbit, Google Fit, Smartwatch, Smart Scale and Fitbit Zip) at different times.

User studies indicated that the core calendar interface enabled users to answer all three classes of questions in Section 7.2.1. It also highlighted some of the challenges we describe below. Figure 7.3 shows data for a hypothetical user, Alice, who had collected physical activity data over 10 years, from 2015-2024. The dark cells indicate the days that she met her 10K steps goal, light cells indicate when she met 50% of her goal and white cells indicate below 50% of her goal. The grey cells are days with no data recorded. The change in cell colour after 2016 reflects when she changed device (from a Fitbit). Superimposed on the figure are icons indicating key challenges in Alice's life and context. Alice started tracking in 2015 as part of a university study where she had a Fitbit Zip. At the time, she was physically active and participated in many student activities. In 2017, she began her working career, entered a serious relationship and became more conscious of her weight gain and lower physical activity. So, she bought her own physical activity tracker along with the Google Fit app and a smart scale. In 2020, her life changed dramatically with marriage and then in 2023 she started a family. From 2019, she stopped using her smartwatch and scale and relied solely on the Google Fit app to keep track of her general health and physical activity.

Interpreting data - Impact of data incompleteness

Alice's data shows the impact of days she did not wear her tracker. This is unsurprising and has been documented among long-term trackers [Tang and Kay, 2017, Epstein et al., 2016b, Meyer et al., 2017]. Figure 7.3 clearly shows the whole days that had no data (the grey cells) but not the impact of wear time (hours per day). For example, if she only wore a tracker for three hours, this may give a white cell - but is three hours enough for the step count to be meaningful so as to answer questions about activity level? User interfaces need to account for this so as to avoid compromising user trust [Tang and Kay, 2017]. Further research is needed to determine how we can ensure users can obtain meaningful and trustworthy answers to their questions, where the information available to the user includes ways to assess the accuracy of the data accounting for incompleteness.

Interpreting data - Scaffolding reflection

While this calendar visualisation proved quite intuitive and effective for answering questions about physical activity, it highlighted the need for scaffolding to help users consider good questions that can be answered with this data. Our study of scaffolding to support reflection [Tang and Kay, 2018a] demonstrated the importance of this scaffolding in helping users consider their goals in reminding users of items of importance, such as weekend versus weekdays activity, that they may not have noticed on their own. Our work has shown the usefulness of our design but also highlights the potential of further studies in providing more personalised and adaptive feedback.

Appendix A

Co-authorship Statements

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Judy Kay**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publications.

Tang, L. M., Day, M., Engelen, L., Poronnik, P., Bauman, A., & Kay, J. (2016). Daily & Hourly Adherence: Towards Understanding Activity Tracker Accuracy. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3211–3218).

Tang, L. M., & Kay, J. (2017). Harnessing Long Term Physical Activity Data—How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 26.

Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 37.

Tang, Lie Ming; Kay, J. (2018). Scaffolding for an OLM for long term physical activity goals. In International Conference on User Modeling, Adaptation, and Personalization.

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscripts, responding to reviewers reports and coordinating submission and publication of the manuscripts.

Signed: Date:

(Judy Kay)

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Adrian Bauman**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publications.

Tang, L. M., Day, M., Engelen, L., Poronnik, P., Bauman, A., & Kay, J. (2016). Daily & Hourly Adherence: Towards Understanding Activity Tracker Accuracy. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3211–3218).

Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 37.

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscripts, responding to reviewers reports and coordinating submission and publication of the manuscripts.

Signed: Date:

(Adrian Bauman)

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Daniel A. Epstein**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publication.

Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 37.

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscript, responding to reviewers reports and coordinating submission and publication of the manuscript.

Signed: Date:

(Daniel A. Epstein)

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Jochen Meyer**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publication.

> Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 37.

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscript, responding to reviewers reports and coordinating submission and publication of the manuscript.

Signed: Date:

(Jochen Meyer)

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Lina Engelen**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publications.

Tang, L. M., Day, M., Engelen, L., Poronnik, P., Bauman, A., & Kay, J. (2016). Daily & Hourly Adherence: Towards Understanding Activity Tracker Accuracy. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3211–3218).

Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 37.

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscripts, responding to reviewers reports and coordinating submission and publication of the manuscripts.

Signed: Date:

(Lina Engelen)

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Philip Poronnik**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publication.

> Tang, L. M., Day, M., Engelen, L., Poronnik, P., Bauman, A., & Kay, J. (2016). Daily & Hourly Adherence: Towards Understanding Activity Tracker Accuracy. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3211–3218).

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscript, responding to reviewers reports and coordinating submission and publication of the manuscript.

Signed: Date:

(Philip Poronnik)

The University of Sydney

Faculty of Engineering and Information Technologies
NSW 2006, Australia

To Whom It May Concern,

I, **Kevin Bragg**, the undersigned are writing this letter to stipulate the role of Lie Ming Tang in the preparation and submission of the following publication.

> Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 37.

Lie Ming Tang during his PhD candidature was responsible for formulating the research questions, conducting study and analyses, writing the draft manuscript, responding to reviewers reports and coordinating submission and publication of the manuscript.

Signed: Date:

(Kevin Bragg)

Appendix B

Human Study Ethics Approval

Research Integrity
Human Research Ethics Committee

Monday, 30 September 2013

Prof Judith Kay
Schl Information Technologies; Faculty of Engineering and Information Technologies
Email: judy.kay@sydney.edu.au

Dear Prof Judith Kay

I am pleased to inform you that the University of Sydney Human Research Ethics Committee (HREC) has approved your project entitled "**Study of metacognitive scaffolding for long term goals.**".

Details of the approval are as follows:

Project No.: **2013/811**

Approval Date: **30 September 2013**

First Annual Report Due: **1 October 2014**

Authorised Personnel: **Kay Judith; Kummerfeld Robert; Tang Lie Ming;**

Documents Approved:

Date Uploaded	Type	Document Name
18/09/2013	Advertisements/Flyer	Advertisement flyer
18/09/2013	Participant Consent Form	Participant Consent Form
18/09/2013	Participant Info Statement	Participant Information statement
26/08/2013	Questionnaires/Surveys	Study questionnaire
21/08/2013	Questionnaires/Surveys	Background questionnaire

HREC approval is valid for four (4) years from the approval date stated in this letter and is granted pending the following conditions being met:

<u>Condition/s of Approval</u>

- Continuing compliance with the National Statement on Ethical Conduct in Research Involving Humans.
- Provision of an annual report on this research to the Human Research Ethics Committee from the approval date and at the completion of the study. Failure to submit reports will result in withdrawal of ethics approval for the project.
- All serious and unexpected adverse events should be reported to the HREC within 72 hours.
- All unforeseen events that might affect continued ethical acceptability of the project should be reported to the HREC as soon as possible.

- Any changes to the project including changes to research personnel must be approved by the HREC before the research project can proceed.

Chief Investigator / Supervisor's responsibilities:

1. You must retain copies of all signed Consent Forms (if applicable) and provide these to the HREC on request.

2. It is your responsibility to provide a copy of this letter to any internal/external granting agencies if requested.

Please do not hesitate to contact Research Integrity (Human Ethics) should you require further information or clarification.

Yours sincerely

Dr Stephen Assinder
Chair
Human Research Ethics Committee

This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (2007), NHMRC and Universities Australia Australian Code for the Responsible Conduct of Research (2007) and the CPMP/ICH Note for Guidance on Good Clinical Practice.

THE UNIVERSITY OF SYDNEY

School of Information Technologies
Faculty of Engineering
and Information Technologies

ABN 15 211 513 464

Professor Judy Kay
Computer Human Adapted Interaction Group
School of Information Technology

Telephone: +61 2 9351 4502
Facsimile: +61 2 9351 3838
Email: judy.kay@sydney.edu.au
Web: www.it.usyd.edu.au/~judy

Room 318,
School of IT, Building J12
University of Sydney,
Sydney NSW 2006,
AUSTRALIA
Web: http://www.sydney.edu.au/

Metacognitive Scaffolding: Towards long term goals!

PARTICIPANT INFORMATION STATEMENT

(1) What is this study about?

You are invited to take part in a research to understand how people view, understand and make use of your own data. Ultimately, the study will provide an initial but important foundation for making data from commercial activity trackers more useful to more people. This may help people achieve their long term goals such as health and well-being.

You have been invited to participate in this study because you have or have used a physical activity tracker or app for more than 1 week. As part of this study, we invite you to an interview with a researcher to share your experiences with your activity tracker. In addition, you have the option of giving us access your physical activity data. We will show you an application with visualisations and reports during the interview and ask you to provide feedback. **Note:** you are still eligible to participate without this.

By participating, you will automatically enter a draw to win a **[prize TBA]** at the end of the study.

This Participant Information Statement tells you about this study. Knowing what is involved will help you decide if you want to take part in the research.

Participation in this research study is entirely VOLUNTARY.

(2) Who is running the study?

Lie Ming Tang will be conducting this study as part of the degree of PhD at the University of Sydney. This will take place under the direct supervision of Professor Judy Kay.

PROFESSOR JUDY KAY | School of Information Technology | Faculty of Engineering
judy.kay@sydney.edu.au
LIE MING TANG | School of Information Technology | Faculty of Engineering
ltan8012@uni.sydney.edu.au

(3) What will the study involve for me?

You will be asked to participate in an interview with the researcher which will take between 30mins to 1 hour. The interview can be via phone, online or in person depending on location, availability and preference. The questions are related to your experiences with the activity tracking devices and how you view, understood and make use of the data. During the interview, we will present an application with visualisation and report of personal activity data and ask you to provide feedback.

We will also collect your activity data (provided that you give permission to do so). The method of collection is dependant on what type of device you have. The researcher will discuss with you the technical details of how your data can be made available to the study based on the type of device you have.

Prior to the start of the interview, you will be asked to fill out a background information sheet. Please note that your personal information will be kept confidential and no identifying information will be used in any publication nor shared third parties. Please see section 9 for details of data collection and privacy.

The interview will be audio recorded for note and transcription purposes only.

(4) How much of my time will the study take?

The background information sheet will take less than 5 mins.
The interview will take between 30mins to 1 hour.

(5) Who can take part in the study?

This interview study is open to any participants who have previously used a personal activity tracking device (e.g., fitbit, Jawbone Up) or app (e.g., Nike+, Endomondo, Google Fit, Apple Health App) for more than 1 weeks.

(6) Do I have to be in the study? Can I withdraw from the study once I've started?

Being in this study is completely voluntary and you do not have to take part. Your decision whether to participate will not affect your current or future relationship with the researchers or anyone else at the University of Sydney.

If you decide to take part in the study and then change your mind later, you are free to withdraw at any time. You can do this by contacting Lie Ming Tang (ltan8012@uni.sydney.edu.au).

Please note that if you do withdraw from the study, you will not be eligible to enter a draw for a prize.

(7) Are there any risks or costs associated with being in the study?

Aside from giving up your time, we do not expect that there will be any risks or costs associated with taking part in this study.

(8) Are there any benefits associated with being in the study?

As part of the interview, we will provide a detailed report on participant's personal activity (provided that data is available). This may provide insights and help users reflect on their ability to achieve their long term goals. However, we cannot guarantee or promise that you will receive any direct benefits from being in the study.

(9) What will happen to information about me that is collected during the study?

By providing your consent, you are agreeing to us collecting your personal activity data (if applicable) as well as interview data as part of our research data.

Your information will be stored securely and your identity/information will be kept strictly confidential, except as required by law. Findings may be published, but you will not be individually identifiable in these publications. Your consent forms will be stored securely in the office of the chief investigator, in a locked cabinet.

The information collected during the interview will be analysed within our group, over time, and between groups. Your personal, individual information will be stored securely and your identity and information will be kept strictly confidential and not disclosed to any other party. Study findings may be published in journals, academic literature or at academic conferences or symposia, but you and your personal data will not be individually identifiable in any of these.

The researchers will hold the data collected during the experiment electronically. The data will be protected by password and will only be accessible by the researchers, except as required by law.

We will keep the information we collect for this study, and we may use it in future projects. By providing your consent you are allowing us to use your information in future projects. We don't know at this stage what these other projects will involve. We will seek ethical approval before using the information in these future projects.

We may give the information from this project to other researchers so that they can use it in their projects. Before we do so, we will take out all the identifying information so that the people we give it to won't know whose information it is. They won't know that you participated in the project and they won't be able to link you to any of the information you provided.

(10) **Can I tell other people about the study?**

Yes, you are welcome to tell other people about the study.

(11) **What if I would like further information about the study?**

When you have read this information, Lie Ming Tang will be available to discuss it with you further and answer any questions you may have. If you would like to know more at any stage during the study, please feel free to contact:
Lie Ming Tang
ltan8012@uni.sydney.edu.au

(12) **Will I be told the results of the study?**

There will be no feedback other than what is provided during the interview in the personal activity report.

(13) **What if I have a complaint or any concerns about the study?**

Research involving humans in Australia is reviewed by an independent group of people called a Human Research Ethics Committee (HREC). The ethical aspects of this study have been approved by the HREC of the University of Sydney [2015/547]. As part of this process, we have agreed to carry out the study according to the *National Statement on Ethical Conduct in Human Research (2007).* This statement has been developed to protect people who agree to take part in research studies.

If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the university using the details outlined below. Please quote the study title and protocol number.

The Manager, Ethics Administration, University of Sydney:

- **Telephone:** +61 2 8627 8176

- **Email:** ro.humanethics@sydney.edu.au
- **Fax:** +61 2 8627 8177 (Facsimile)

This information sheet is for you to keep

Bibliography

[Ahmad and Bull, 2009] Ahmad, N. and Bull, S. (2009). Learner Trust in Learner Model Externalisations. In *AIED*, pages 617–619.

[Al-Shanfari et al., 2016] Al-Shanfari, L., Epp, C. D., and Bull, S. (2016). Uncertainty in Open Learner Models: Visualising Inconsistencies in the Underlying Data. In *LAL@ LAK*, pages 23–30.

[Albert, 2017] Albert, L. (2017). The Surprising Potential Fitness Tracker Buyer.

[Althoff, 2017] Althoff, T. (2017). Population-Scale Pervasive Health. *IEEE Pervasive Computing*, 16(4):75–79.

[Althoff et al., 2017] Althoff, T., Sosič, R., Hicks, J. L., King, A. C., Delp, S. L., and Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*.

[a.W.M.M. Aleven and Koedinger, 2002] a.W.M.M. Aleven, V. and Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26(2):147–179.

[Azevedo and Cromley, 2004] Azevedo, R. and Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of educational psychology*, 96(3):523.

[Azevedo et al., 2010] Azevedo, R., Johnson, A., Chauncey, A., and Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In *New science of learning*, pages 225–247. Springer.

[Bandura, 2005a] Bandura, A. (2005a). The growing centrality of self-regulation in health promotion and disease prevention. *The european health psychologist*, 1:11–12.

[Bandura, 2005b] Bandura, A. (2005b). The Primacy of Self-Regulation in Health Promotion. *Applied Psychology*, 54(2):245–254.

[Barua, 2016] Barua, D. (2016). A time to remember , a time to forget : Enabling people to control long term sensor data .

[Barua et al., 2014] Barua, D., Kay, J., Kummerfeld, B., and Paris, C. (2014). Modelling long term goals. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 1–12. Springer.

[Barua et al., 2013] Barua, D., Kay, J., and Paris, C. (2013). Viewing and Controlling Personal Sensor Data: What Do Users Want? In Berkovsky, S. and Freyne, J., editors, *Persuasive Technology*, volume 7822 of *Lecture Notes in Computer Science*, pages 15–26. Springer Berlin Heidelberg.

[Behrens and Dinger, 2005] Behrens, T. K. and Dinger, M. K. (2005). Ambulatory Physical Activity Patterns of College Students. *American Journal of Health Education*, 36(4):221–227.

[Bentley et al., 2013] Bentley, F., Tollmar, K., and Stephenson, P. (2013). Health Mashups: Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. *Tochi*, 20(5):1–27.

[Bragg, 2015] Bragg, K. A. (2015). *Does The Quantified Self Equal Quantified Health?* PhD thesis, University of Sydney.

[Bravata et al., 2007] Bravata, D. M., Smith-Spangler, C., Sundaram, V., Gienger, A. L., Lin, N., Lewis, R., Stave, C. D., Olkin, I., and Sirard, J. R. (2007). Using pedometers to increase physical activity and improve health: a systematic review. *JAMA*, 298(19):2296–2304.

[Buckworth, 2012] Buckworth, J. (2012). Exercise Adherence in College Students: Issues and Preliminary Results. *Quest*, 53(3):335–345.

[Bull et al., 1995] Bull, S., Brna, P., and Pain, H. (1995). Extending the scope of the student model. *User Modeling and User-Adapted Interaction*, 5(1):45–65.

[Bull et al., 2016] Bull, S., Ginon, B., Boscolo, C., and Johnson, M. (2016). Introduction of Learning Visualisations and Metacognitive Support in a Persuadable Open Learner Model. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 30–39, New York, NY, USA. ACM.

[Bull and Kay, 2010] Bull, S. and Kay, J. (2010). Open learner models. In *Advances in intelligent tutoring systems*, pages 301–322. Springer.

[Bull and Kay, 2013] Bull, S. and Kay, J. (2013). Open learner models as drivers for metacognitive processes. In *International handbook of metacognition and learning technologies*, pages 349–365. Springer.

[Bull and Kay, 2016] Bull, S. and Kay, J. (2016). SMILI: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, 26(1):293–331.

[Cadmus-Bertram et al., 2015a] Cadmus-Bertram, L., Marcus, B. H., Patterson, R. E., Parker, B. A., and Morey, B. L. (2015a). Use of the Fitbit to Measure Adherence to a Physical Activity Intervention Among Overweight or Obese, Postmenopausal Women: Self-Monitoring Trajectory During 16 Weeks. *JMIR mHealth and uHealth*, 3(4):e96.

[Cadmus-Bertram et al., 2015b] Cadmus-Bertram, L. A., Marcus, B. H., Patterson, R. E., Parker, B. A., and Morey, B. L. (2015b). Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine*.

[Choe et al., 2017] Choe, E. K., Lee, B., Zhu, H., Riche, N. H., and Baur, D. (2017). Understanding Self-Reflection: How People Reflect on Personal Data through Visual Data Exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'17). ACM, New York, NY, USA*, volume 10.

[Choe et al., 2014] Choe, E. K., Lee, N. B., Lee, B., Pratt, W., and Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1143–1152. ACM.

[Church and Blair, 2009] Church, T. S. and Blair, S. N. (2009). When will we treat physical activity as a legitimate medical therapy...even though it does not come in a pill? *British journal of sports medicine*, 43(2):80–81.

[Clawson et al., 2015] Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D., and Mamykina, L. (2015). No Longer Wearing : Investigating the Abandonment of Personal Health-Tracking Technologies on Craigslist. *Ubicomp 2015*.

[Consolvo et al., 2009] Consolvo, S., Klasnja, P., McDonald, D. W., and Landay, J. A. (2009). Goal-setting considerations for persuasive technologies that encourage physical activity. In *Proceedings of the 4th International Conference on Persuasive Technology - Persuasive '09*, page 1, New York, New York, USA. ACM Press.

[Consolvo et al., 2014] Consolvo, S., Klasnja, P., McDonald, D. W., Landay, J. A., et al. (2014). Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness. *Foundations and Trends® in Human–Computer Interaction*, 6(3–4):167–315.

[Consolvo et al., 2008] Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., and Others (2008). Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1797–1806. ACM.

[Daskalova et al., 2017] Daskalova, N., Desingh, K., Papoutsaki, A., Schulze, D., Sha, H., and Huang, J. (2017). Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):46.

[Deci and Ryan, 2000] Deci, E. L. and Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4):227–268.

[Dennison et al., 2013] Dennison, L., Morrison, L., Conway, G., and Yardley, L. (2013). Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. *Journal of medical Internet research*, 15(4):e86.

[Desharnais et al., 1986] Desharnais, R., Bouillon, J., and Godin, G. (1986). Self-efficacy and outcome expectations as determinants of exercise adherence. *Psychological Reports*, 59(3):1155–1159.

[Desmarais and Baker, 2012] Desmarais, M. C. and Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38.

[Doherty et al., 2017] Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van Hees, V. T., Trenell, M. I., Owen, C. G., and Others (2017). Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank Study. *PloS one*, 12(2):e0169649.

[Duffy and Azevedo, 2015] Duffy, M. C. and Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior*, 52:338–348.

[Elsden et al., 2015] Elsden, C., Kirk, D. S., and Durrant, A. C. (2015). A Quantified Past: Toward Design for Remembering With Personal Informatics. *Human–Computer Interaction*, pages 1–40.

[Endeavour, 2014] Endeavour (2014). Inside wearables part 2, How the Science of Behavior Change Offers the Secret to Long-Term Engagement. (June).

[Epp and Bull, 2015] Epp, C. D. and Bull, S. (2015). Uncertainty representation in visualizations of learning analytics for learners: current approaches and opportunities. *IEEE Transactions on Learning Technologies*, 8(3):242–260.

[Epstein et al., 2014] Epstein, D., Cordeiro, F., Bales, E., Fogarty, J., and Munson, S. (2014). Taming data complexity in lifelogs: exploring visual cuts of personal informatics data. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 667–676. ACM.

[Epstein, 2015] Epstein, D. A. (2015). Personal informatics in everyday life. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp '15*, (Figure 1):429–434.

[Epstein et al., 2016a] Epstein, D. A., Caraway, M., Johnston, C., Ping, A., Fogarty, J., and Munson, S. A. (2016a). Beyond Abandonment to Next Steps: Understanding and Designing for Life After Personal Informatics Tool Use. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1109–1113.

[Epstein et al., 2016b] Epstein, D. A., Kang, J., Pina, L. R., Fogarty, J., and Munson, S. A. (2016b). Reconsidering the Device in the Drawer: Lapses as a Design Opportunity in Personal Informatics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 829—-840. ACM.

[Evenson et al., 2015] Evenson, K. R., Goto, M. M., and Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *The international journal of behavioral nutrition and physical activity*, 12(1):159.

[Fairclough et al., 2014] Fairclough, S. J., Boddy, L. M., Mackintosh, K. a., Valencia-Peris, A., and Ramirez-Rico, E. (2014). Weekday and weekend sedentary time and physical activity in differentially active children. *Journal of Science and Medicine in Sport*, In Press.

[Fan et al., 2012] Fan, C., Forlizzi, J., and Dey, A. K. (2012). A spark of activity: exploring informative art as visualization for physical activity. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 81–84, New York, NY, USA. ACM.

[Faust et al., 2017] Faust, L., Purta, R., Hachen, D., Striegel, A., Poellabauer, C., Lizardo, O., and Chawla, N. V. (2017). Exploring Compliance: Observations from a Large Scale Fitbit Study. In *Proceedings of the 2nd International Workshop on Social Sensing*, pages 55–60. ACM.

[Ferreira et al., 2007] Ferreira, I., Van Der Horst, K., Wendel-Vos, W., Kremers, S., Van Lenthe, F. J., and Brug, J. (2007). Environmental correlates of physical activity in youth - A review and update. *Obesity Reviews*, 8(2):129–154.

[Feyzi-Behnagh et al., 2014] Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., and Crowley, R. S. (2014). Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional science*, 42(2):159–181.

[Fogg, 2003] Fogg, B. J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*, volume 5 of *The Morgan Kaufmann series in interactive technologies*. Morgan Kaufmann.

[Fritz et al., 2014] Fritz, T., Huang, E. M., Murphy, G. C., and Zimmermann, T. (2014). Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 487–496. ACM.

[Gouveia et al., 2015] Gouveia, R., Karapanos, E., and Hassenzahl, M. (2015). How do we engage with activity trackers?: a longitudinal study of habito. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1305–1316. ACM.

[Guerra, 2016] Guerra, J. (2016). Open Social Learner Models for Self-Regulated Learning and Learning Motivation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 329–332. ACM.

[Guerra-Hollstein et al., 2017] Guerra-Hollstein, J., Barria-Pineda, J., Schunn, C. D., Bull, S., and Brusilovsky, P. (2017). Fine-Grained Open Learner Models: Complexity Versus Support. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 41–49. ACM.

[Harrison et al., 2015] Harrison, D., Marshall, P., Bianchi-berthouze, N., and Bird, J. (2015). Activity Tracking Barriers, Workarounds, and Customisation. *Ubicomp 2015*, page 617.

[Haskell et al., 2007] Haskell, W. L., Lee, I., Pate, R. R., Powell, K. E., Blair, S. N., Franklin, B. A., Macera, C. A., Heath, G. W., Thompson, P. D., Bauman, A., and Others (2007). Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Medicine and science in sports and exercise*, 39(8):1423.

[Huang, 2016] Huang, D. (2016). *Visualizing personal data in context: an on-calendar design strategy for behaviour feedback*. PhD thesis, University of Victoria.

[IDC, 2017] IDC (2017). Worldwide Quarterly Wearable Device Tracker.

[Jeong et al., 2017] Jeong, H., Kim, H., Kim, R., Lee, U., and Jeong, Y. (2017). Smartwatch Wearing Behavior Analysis : A Longitudinal Study. *ACM Ubicomp 2017 / Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1(3).

[Kay, 1994] Kay, J. (1994). The um toolkit for cooperative user modelling. *User Modeling and User-Adapted Interaction*, 4(3):149–196.

[Kay, 2016] Kay, J. (2016). Enabling people to harness and control EDM for lifelong, life-wide learning. In *EDM*, pages 10–20.

[Kay and Kummerfeld, 2012] Kay, J. and Kummerfeld, B. (2012). Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *TiiS*, 2(4):24.

[Kay et al., 2016] Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. (2016). When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5092–5103. ACM.

[Keating et al., 2015] Keating, X. D., Guan, J., Piñero, J. C., and Bridges, D. M. (2015). A meta-analysis of college students' physical activity behaviors. *Journal of American college health : J of ACH*, 54(2):116–125.

[Konstabel et al., 2014] Konstabel, K., Veidebaum, T., Verbestel, V., Moreno, L. A., Bammann, K., Tornaritis, M., Eiben, G., Molnar, D., Siani, A., Sprengeler, O., Wirsik, N., Ahrens, W., and Pitsiladis, Y. (2014). Objectively measured physical activity in European children: the IDEFICS study. *Int J Obes (Lond)*, 38 Suppl 2(S2):S135–43.

[Kuo et al., 2017] Kuo, T. T., Kim, H. E., and Ohno-Machado, L. (2017). Blockchain distributed ledger technologies for biomedical and health care applications. *Journal of the American Medical Informatics Association*, 24(6):1211–1220.

[Lazar et al., 2015] Lazar, A., Tanenbaum, J., Koehler, C., and Nguyen, D. H. (2015). Why We Use and Abandon Smart Devices. *Ubicomp 2015*, page 635.

[Li et al., 2010] Li, I., Dey, A., and Forlizzi, J. (2010). A stage-based model of personal informatics systems. *Proceedings of the 28th international conference on Human factors in computing systems CHI 10*, (August 2015):557.

[Li et al., 2011] Li, I., Dey, A. K., and Forlizzi, J. (2011). Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, pages 405–414, New York, NY, USA. ACM.

[Li et al., 2012] Li, I., Dey, A. K., and Forlizzi, J. (2012). Using context to reveal factors that affect physical activity. *ACM Transactions on Computer-Human Interaction*, 19(1):1–21.

[Lin et al., 2006] Lin, J., Mamykina, L., Lindtner, S., Delajoux, G., and Strub, H. (2006). Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game. In Dourish, P. and Friday, A., editors, *UbiComp 2006: Ubiquitous Computing*, volume 4206 of *Lecture Notes in Computer Science*, pages 261–278. Springer Berlin Heidelberg.

[Liou et al., 2016] Liou, K., Ho, S., Fildes, J., and Ooi, S.-Y. (2016). High intensity interval versus moderate intensity continuous training in patients with coronary artery disease: a meta-analysis of physiological and clinical parameters. *Heart, Lung and Circulation*, 25(2):166–174.

[Liu et al., 2011] Liu, Y., Gummadi, K. P., Krishnamurthy, B., and Mislove, A. (2011). Analyzing Facebook Privacy Settings: User Expectations vs. Reality. *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pages 61–70.

[Locke and Latham, 2002a] Locke, E. A. and Latham, G. P. (2002a). Building a Practically Useful Theory of Goal Setting and Task Motivation – A 35-Year Odyssey.

[Locke and Latham, 2002b] Locke, E. A. and Latham, G. P. (2002b). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705.

[Long and Aleven, 2011] Long, Y. and Aleven, V. (2011). Students' understanding of their student model. In *International Conference on Artificial Intelligence in Education*, pages 179–186. Springer.

[Long and Aleven, 2013] Long, Y. and Aleven, V. (2013). Supporting students' self-regulated learning with an open learner model in a linear equation tutor. In *Artificial intelligence in education*, pages 219–228. Springer.

[Marcengo et al., 2016] Marcengo, A., Rapp, A., Cena, F., and Geymonat, M. (2016). The Falsified Self: Complexities in Personal Data Collection. *UAHCI 2016*, 9737:351–358.

[Marcus et al., 1992] Marcus, B. H., Selby, V. C., Niaura, R. S., and Rossi, J. S. (1992). Self-efficacy and the stages of exercise behavior change. *Research quarterly for exercise and sport*, 63(1):60–66.

[Matthews et al., 2008] Matthews, C. E., Chen, K. Y., Freedson, P. S., Buchowski, M. S., Beech, B. M., Pate, R. R., and Troiano, R. P. (2008). Amount of time spent in sedentary behaviors in the United States, 2003-2004. *American Journal of Epidemiology*, 167(7):875–881.

[Meyer et al., 2016a] Meyer, J., Heuten, W., and Boll, S. (2016a). No Effects But Useful ? Long Term Use of Smart Health Devices. *Ubicomp/ISWC'16 Adjunct*, pages 516–521.

[Meyer et al., 2016b] Meyer, J., Schnauber, J., Heuten, W., Wienbergen, H., Hambrecht, R., Appelrath, H.-J., and Boll, S. (2016b). Exploring Longitudinal Use of Activity Trackers. *Procedings of IEEE ICHI - International Conference on Healthcare Informatics*, pages 198–206.

[Meyer et al., 2017] Meyer, J., Wasmann, M., Heuten, W., El Ali, A., and Boll, S. (2017). Identification and Classification of Usage Patterns in Long-Term Activity Tracking. *CHI '17 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[Migueles et al., 2017] Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nystr??m, C., Mora-Gonzalez, J., L??f, M., Labayen, I., Ruiz, J. R., and Ortega, F. B. (2017). Accelerometer Data Collection and Processing Criteria to Assess

Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Medicine*, pages 1–25.

[Mitrovic and Martin, 2002] Mitrovic, A. and Martin, B. (2002). Evaluating the effects of open student models on learning. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 296–305. Springer.

[Mitrovic and Martin, 2007] Mitrovic, A. and Martin, B. (2007). Evaluating the effect of open student models on self-assessment. *International Journal of Artificial Intelligence in Education*, 17(2):121–144.

[Nielsen, 1994] Nielsen, J. (1994). *Usability engineering*. Elsevier.

[Rapp and Cena, 2016] Rapp, A. and Cena, F. (2016). Personal Informatics for Everyday Life: How Users without Prior Self-Tracking Experience Engage with Personal Data. *International Journal of Human-Computer Studies*, 94:1–17.

[Rooksby et al., 2014] Rooksby, J., Rost, M., Morrison, A., and Chalmers, M. C. (2014). Personal tracking as lived informatics. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1163–1172. ACM.

[Rooney et al., 2003] Rooney, B., Smalley, K., Larson, J., and Havens, S. (2003). Is knowing enough? Increasing physical activity by wearing a pedometer. *WMJ-MADISON-*, 102(4):31–36.

[Saelens et al., 2012] Saelens, B. E., Sallis, J. F., Frank, L. D., Couch, S. C., Zhou, C., Colburn, T., Cain, K. L., Chapman, J., and Glanz, K. (2012). Obesogenic neighborhood environments, child and parent obesity: The neighborhood impact on kids study. *American Journal of Preventive Medicine*, 42(5):e57—-e64.

[Shih et al., 2015] Shih, P. C., Han, K., Poole, E. S., Rosson, M. B., and Carroll, J. M. (2015). Use and adoption challenges of wearable activity trackers. *iConference 2015 Proceedings*.

[Shilts et al., 2004] Shilts, M. K., Horowitz, M., and Townsend, M. S. (2004). Goal setting as a strategy for dietary and physical activity behavior change: a review of the literature. *American journal of health promotion : AJHP*, 19(2):81–93.

[Shneiderman, 1996] Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for informatio nvisualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343.

[Sisson et al., 2015] Sisson, S. B., McClain, J. J., and Tudor-Locke, C. (2015). Campus walkability, pedometer-determined steps, and moderate-to-vigorous physical activity: a comparison of 2 university campuses. *Journal of American college health : J of ACH*, 56(5):585–592.

[Strecher et al., 1995] Strecher, V. J., Seijts, G. H., Kok, G. J., Latham, G. P., Glasgow, R., DeVellis, B., Meertens, R. M., and Bulger, D. W. (1995). Goal setting as a strategy for health behavior change. *Health education quarterly*, 22(2):190–200.

[Tabuenca et al., 2015] Tabuenca, B., Kalz, M., Drachsler, H., and Specht, M. (2015). Time will tell: The role of mobile learning analytics in self-regulated learning. *Computers & Education*, 89:53–74.

[Tang and Kay, 2016] Tang, L. M. and Kay, J. (2016). Daily & hourly adherence : towards understanding activity tracker accuracy. *CHI '16 Extended Abstracts on Human Factors in Computing Systems.*

[Tang and Kay, 2017] Tang, L. M. and Kay, J. (2017). Harnessing Long Term Physical Activity Data—How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):26.

[Tang and Kay, 2018a] Tang, L. M. and Kay, J. (2018a). Scaffolding for an olm for long-term physical activity goals. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 147–156. ACM.

[Tang and Kay, 2018b] Tang, L. M. and Kay, J. (2018b). Understanding physical activity tracking data: wear-time matters. *Pending Submission.*

[Tang et al., 2018] Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., and Kay, J. (2018). Defining adherence : making sense of physical activity tracker data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):37.

[Tiedemann et al., 2015] Tiedemann, A., Hassett, L., and Sherrington, C. (2015). A novel approach to the issue of physical inactivity in older age. *Preventive medicine reports*, 2:595–597.

[Togo et al., 2008] Togo, F., Watanabe, E., Park, H., Yasunaga, A., Park, S., Shephard, R. J., and Aoyagi, Y. (2008). How many days of pedometer use predict the annual activity of the elderly Reliably? *Medicine and Science in Sports and Exercise*, 40(6):1058–1064.

[Trost et al., 2005] Trost, S. G., Mciver, K. L., and Pate, R. R. (2005). Conducting accelerometer-based activity assessments in field-based research. *Medicine and Science in Sports and Exercise*, 37(11 SUPPL.):531–543.

[Tucker et al., 2011] Tucker, J. M., Welk, G. J., and Beyler, N. K. (2011). Physical activity in U.S. adults: Compliance with the physical activity guidelines for Americans. *American Journal of Preventive Medicine*, 40(4):454–461.

[Tudor-locke, 2016] Tudor-locke, C. (2016). The Objective Monitoring of Physical Activity: Contributions of Accelerometry to Epidemiology, Exercise Science and Rehabilitation.

[Tudor-Locke et al., 2015] Tudor-Locke, C., Barreira, T. V., Schuna, J. M., Mire, E. F., Chaput, J.-P., Fogelholm, M., Hu, G., Kuriyan, R., Kurpad, A., Lambert, E. V., and Others (2015). Improving wear time compliance with a 24-hour waist-worn accelerometer protocol in the International Study of Childhood Obesity, Lifestyle and the Environment (ISCOLE). *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):11.

[Tudor-Locke et al., 2004] Tudor-Locke, C., Bassett, D. R., Swartz, A. M., Strath, S. J., Parr, B. B., Reis, J. P., Dubose, K. D., and Ainsworth, B. E. (2004). A preliminary study of one year of pedometer self-monitoring. *Ann Behav Med*, 28.

[Tudor-Locke et al., 2011] Tudor-Locke, C., Craig, C. L., Brown, W. J., Clemes, S. A., De Cocker, K., Giles-Corti, B., Hatano, Y., Inoue, S., Matsudo, S. M., Mutrie, N., Oppert, J.-M., Rowe, D. A., Schmidt, M. D., Schofield, G. M., Spence, J. C., Teixeira, P. J., Tully, M. A., and Blair, S. N. (2011). How many steps/day are enough? for adults. *International Journal of Behavioral Nutrition and Physical Activity*, 8(1):79.

[Tudor-Locke et al., 2008] Tudor-Locke, C., Hatano, Y., Pangrazi, R. P., and Kang, M. (2008). Revisiting "How many steps are enough?". *Med Sci Sports Exerc*, 40.

[Yang et al., 2015] Yang, R., Shin, E., Newman, M. W., and Ackerman, M. S. (2015). When Fitness Trackers Don't 'Fit': End-User Difficulties in the Assessment of Personal Tracking Device Accuracy. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.