Working Paper Series
ISSN 1177-777X

INVESTIGATING THE USE OF ACTIVITY TRACKERS TO OBSERVE HIGH-RISK WORK ENVIRONMENTS

**Judy Bowen, Annika Hinze,
Sally Jo Cunningham and Richard Parker**

Working Paper: 01/2015
November 2015

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

This working paper is the extended version of the publication:
J. Bowen, A. Hinze, S.J. Cunningham, Richard Parker (2015) "Evaluating Low-cost Activity Trackers for Use in Large-scale Data Gathering of Forestry Workers" in Proc. of the OzCHI Australian Conference on Human-Computer Interaction, December 2015.

Investigating the use of Activity Trackers to Observe High-Risk Work Environments

Judy Bowen,[1] Annika Hinze,[1] Sally Jo Cunningham,[1] Richard Parker[2]
[1]The University of Waikato, New Zealand,
{jbowen,hinze,sallyjo}@waikato.ac.nz
[2] Scion Research, Christchurch, New Zealand,
Richard.Parker@scionresearch.com

Abstract

The New Zealand forestry industry has the country's highest rate of workplace fatalities. The reasons are not well studied or understood and no large-scale systematic physical and physiological data has been recorded to investigate this. Current research focusses on developing mechanised solutions and changing worker behaviour. We believe the first step in identifying any successful solution is to develop a fine-grained understanding of the physical context of forestry work by performing large-scale data collection of the levels of physical activity the workers engage in as well as their sleep patterns over extended periods of time. Our goal is to use lightweight, wearable technology (so-called activity trackers) to collect this data. In order to do so we need a clear understanding of the capabilities and limitations of such devices, both in general and in the proposed use environment for forestry workers. In this paper we present the results of user studies and comparisons of six activity trackers and three mobile phone applications used to track activity and sleep. We also discuss our initial pilot study with forestry workers and discuss the problems encountered using the trackers in the environment.

1 Introduction

Forestry work in New Zealand is largely done manually, due to the topography of the land which is extremely hilly at best and mountainous at worst.

Even though the workers operate in teams, while operating chainsaws they need to be relatively far away from each other for safety. Thus work is often done in solitude,[1] after (sometimes) long hours spent driving out to the forest in the early morning. Despite world-wide advances in forestry safety, the New Zealand statistics continue to

[1]Even though the workers operate in teams, while they are operating chainsaws they need to be relatively far away from each other for safety.

Figure 1: Typical NZ forestry terrain

worsen [1]. The NZ forestry industry has the country's highest rate of workplace fatalities[2] with 10 in 2013 [2]. It also has the highest rate of workplace injury in the country, with numbers increasing annually. 188 serious harm notifications were recorded in 2012 (879 since 2008), and the rate of accident compensation claims for the forestry sector is almost six times that for all other sectors [3]. Different groups involved in the forestry industry (from government ministries to the workers themselves) have differing opinions on the causes of the high accident rate, but no large scale data exists in New Zealand on the working conditions of forestry workers. In order to begin investigating suitable ways to address the unacceptable accident rate, a large-scale data gathering exercise needs to take place as a first step to aid identification of the key factors and relevant aspects of the workers' activities. We also believe that a wider context needs to be taken into account beyond the time spent during working hours in the forestry environment: e.g., the number of hours spent driving to get to and from work sites, early waking hours required to ensure workers arrive in remote sites on time, the high levels of physical exertion the work demands etc. To collect and analyse large quantities of data in a non-intrusive way, light-weight and affordable devices are needed. These need to be unobtrusive to the wearers as they work and so also need to be low-cost to prevent them becoming a burden to the forestry workers, who may otherwise feel they have to look after the equipment and make sure it doesn't get damaged or lost. Our project, therefore, explores the use of cheap, wearable activity tracking devices originally designed for personal fitness regimes. This paper reports on an initial exploration of the suitability of using activity trackers for recording activity and sleep data of large numbers of forestry workers over an extended period of time. The contributions provided by this paper are:

- a comparison of six activity trackers and three phone apps for step counting and sleep evaluation based on user studies
- calibration of movement detection for several trackers
- identification of usability considerations when using these devices over extended periods of time (both in forestry and general environments)
- a discussion of the use of activity trackers for forestry safety monitoring

The remainder of the paper is organised as follows: Section 2 provides some more background to the forestry industry and a discussion of the limited existing research. Section 3 then introduces the activity trackers and mobile phone applications described in this paper and discusses the relevance of the data they can provide with respect

[2]The annual industry average in New Zealand is 5 fatalities per year [2]

to well-known principles of fatigue, performance and safety. Section 4 presents the results of our calibration and longer-term user studies and introduces our ongoing pilot study with forestry workers. Finally we discuss the outcomes from the work so far and present our conclusions and plans for future work.

2 Background: Forestry Safety

Forestry is labour-intensive: average working hours are 40-60 per week and many tasks involve high levels of physical activity (estimated as the equivalent of running a half marathon each working day [4]). An independent safety review has recently been conducted based on surveys and interviews with employees and workers across all sectors of the industry [1]. The issue of worker fatigue, long working hours, physically demanding tasks and potential for poor quality sleep are all identified as common themes in areas likely to contribute to poor safety [2, 3, 1]. Additionally, the difficulty in identifying and monitoring these factors was likewise identified. However, no systematic data has ever been recorded to test these hypotheses and major stakeholders (such as worker unions, government agencies etc.) have conflicting views of the most significant contributing factors to accidents. One of the few forestry-related studies involved the use of video-cameras being worn by eight workers to try and gain insight into the differing working practices of novices vs. experienced tree fellers [5]. In initial stages of this project Parker attempted to record workers himself as an observer in the work areas, however it became apparent that these types of observational studies by researchers are not possible in such high-risk environments as what was observed was the workers trying to make sure the researcher stayed safe and out of harm's way rather than their usual work practices. As a result of this Parker developed a light-weight, wearable camera to be used by the observed forestry workers. Although some valuable insights were gained, the study was necessarily limited by the number of workers who could be fitted with the cameras and the length of time that the cameras could be worn.

Because of the logistic issues in gaining access to gather accurate data, much of the current work in forestry safety improvements focusses on developing mechanised solutions which remove workers from the equation wherever possible, or on introducing regulations and compliance requirements to try to modify worker behaviours. More recently, the use of monitoring vests (similar to those worn by athletes) was proposed. These vests track heart-rate, cardiovascular intensity, hydration levels and exertions [4]. While these may provide useful insights and in-situ data which may benefit the longer term study, they are not suitable for the large-scale data gathering we propose (due to cost and the fact that they are not unobtrusive) and do not enable us to monitor workers over 24 hour periods (one of the crucial elements of our study).

3 Background: Project context

Our goal is to find practical ways to gather long-term, real-world data from forestry workers as a starting point to identifying potential hazard situations and finding ways of avoiding them. While the continuing emergence of hi-tech solutions (particularly in areas such as video and voice capture) can be used to provide unobtrusive ways of capturing information, they can still be problematic for large-scale data-gathering experiments, particularly in remote environments (due to cost and deployment issues). To this end we propose the use of light-weight, low-cost, wearable activity trackers as a practical tool for gathering such data. We want to collect data not just during working hours, but 24 hours a day, and consider both the activity levels and the sleep quality of the workers. This approach will allow us to look at some of the anecdotal contributing factors to the high accident rate, for example: ‘‘.. workers turn up on Monday morning having played rugby and partied hard all weekend, they're exhausted before they start" or the fact that workplaces are remote, meaning many workers have to get up early for a long drive before they begin work; as well as to consider factors such as poor sleep, or lack of sleep in general and the effect this might have over an extended period of time. The activity trackers are ideal for this purpose as they are designed to be worn both during the day and night and to monitor exactly this type of data.

Our project consists of five stages, the first of which – determining which activity trackers are the most appropriate to be used for forestry workers and how best to use them – we report in this paper. The subsequent stages will involve: large-scale data collection; data analysis and identification of patterns that might be consistent with work place safety incidents; development of models of safety properties and scenarios; implementation of devices with feedback, which use captured data and model patterns to identify potentially dangerous conditions as they emerge. In this paper we discuss our investigation and comparison of various activity trackers to identify their suitability (or not) for use with forestry workers. We also discuss our pilot studies, which use selected devices for small-scale data gathering exercises with forestry workers.

4 Sleep and Activity Tracking

The growth in popularity of personal activity trackers has seen a similar growth in research interest in these devices. This ranges from consumer-type comparisons, which aim to find the 'best' device for a particular use or demographic, to more serious scientific studies investigating accuracy of the devices for specific tasks [6] or activity/sleep monitoring in general [7, 8]. The "gold standard" for sleep quality measurement is polysomnography. This involves an in-laboratory study of one or more night's sleep with the participant having a variety of different sensors attached to their body, and typically includes video and sound monitoring. Polysomnography comes at a cost – notably the intrusive nature of the monitoring and the cost in terms of both equipment and time. This means that it is typically not a suitable solution for large-scale sleep monitoring as proposed here.

[9] described the challenges of activity recognition using on-body sensors such as activity trackers. A number of recent health studies tested the validity of vendor claims about activity trackers [10, 11]. Prince et al (2008) found that self-reported and directly measured physical activities had varying degrees of correlation; a similar result to the one found in [12] and [13].

In [6] a comparison is made between the Fitbit activity tracker's sleep tracking and polysomnography in an attempt to quantify how accurate the Fitbit is. The findings show that the device is suitable for sleep tracking of the general population (although there is tendency for both sleep time and quality to be over-estimated) but is not appropriate for determining or diagnosing sleep disorders. There are also higher-tech sleep tracking solutions such as the Readiband (produced by FatigueScience) used with high-performance athletes and military personnel which are seen as the next best alternative to polysomnography. These are wrist-worn devices similar in appearance to a smart watch, however, the cost of these devices (and also supporting software costs) mean that they are not suitable for our studies with forestry workers. Our investigations fo-

Figure 2: Polar Loop, Fitbit Flex, Fitbit One, Pebble, Jawbone UP and Withings Pulse

cus on the both the general accuracy and compatibility between devices as well as the specific requirements for use as activity and sleep trackers for forestry workers. We present details of our studies next.

5 Comparison of trackers

Numerous different activity trackers and tracking applications for mobile devices exist, ranging from simple pedometer-style devices to fully featured smart-watches which integrate with phone applications. New ones come to the market almost weekly, and the range of functionalities continue to grow and improve. Our first step was to filter the large number of existing devices down to a manageable subset (initially by excluding those not available to the NZ market or over a particular price limit) and then take those that appeared to be suitable for our requirements and using information provided by their manufacturers, categorise them to determine whether or not they meet certain criteria. Table 1 provides an overview of some of this information.

Device and app selection. We are most interested in small, light-weight, low-cost devices (less than NZ$200 or US$160) which have at least the functionality to track the number of steps taken per day, as well as perform sleep analysis. We discuss in this paper six devices – Fitbit Flex, Fitbit One, Jawbone UP, Polar Loop, Withings Pulse and Pebble smart watch, see Figure 2 – and three mobile applications – SleepAs (Android) and Sleep Cycle (iOS) for sleep monitoring, and S-Health (Samsung) for step counting.

Device/app format and functionality. The Fitbit Flex, Jawbone UP, and Polar loop are all worn on the wrist and are slightly smaller than an average wristwatch; the Pebble is a wrist-worn smart watch; the Fitbit One and Withings Pulse are designed to be carried in a pocket or clipped onto a belt loop, and are then transferred to a wristband for sleep monitoring. All of these devices have the same basic functionality: they record steps taken by the wearer and perform sleep monitoring. Sleep monitoring includes length of time slept, number of times woken during the night, depth of sleep (from deep to light as well as identifying restless sleep) and an overall sleep quality score. Some devices also detect elevation (in terms of stair climbing) and an estimate of calories burned (see Table 1). All of the calculations are based on motion tracking, but each device uses different algorithms to achieve this (for more details see next section). The mobile sleep apps SleepAs and SleepCycle require the user to launch the application on their phone, start the sleep mode and then place the phone in a specified position (typically next to the pillow, sometimes under the mattress cover to ensure it stays in place). Some of the apps recommend that the phone is also connected to a power supply during the night (to ensure battery does not run out). The step counting app, S-Health, needs to be launched once and then runs continuously in the background.

Data presentation Three of the devices (Fitbit One, Withings Pulse and Polar Loop) have displays, which provide quick access to data to the user. For all devices full data, with varying levels of analysis, are provided by way of mobile phone apps and web

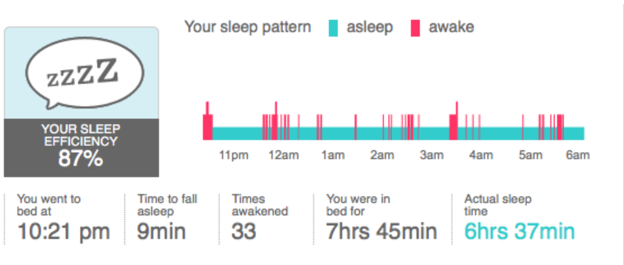

Figure 3: Fitbit online view

Device	How Worn	Steps	Distance	Sleep	Stairs	Calories	GPS	Water Resistent	Battery	Price
FitBit One	Clips to clothes	Yes	Yes	Yes	Yes	Yes	No	Yes	Rechargable	NZ $160
FitBit Flex	Wristband	Yes	Yes	Yes	No	Yes	No	Yes	Rechargable	NZ $160
FitBit Zip	Clips to clothes	Yes	Yes	No	No	Yes	No	Yes	Replace	NZ $100
Galaxy Gear	Wristband	Yes	Yes	No	No	with app	Yes	?	Replace	NZ $450
Pebble	Wristband	Yes	Yes	With app	No	with app	Yes	Yes	Rechargable	US $250
Sony Smartwatch	Wristband	Yes	Yes	With app	?	?	?	No	Rechargable	NZ $300
Nike Fuelband	Wristband	Yes	Yes	No	?	Yes	?	?	Rechargable	US $150
Misfit Wearables	Various	Yes	Yes	Yes	?	Yes	?	Yes	Replace	US$170
Polar Loop	Wristband	Yes	Yes	Yes	?	Yes	No	Yes	Rechargable	NZ $150
Withings Pulse	Clips to clothes	Yes	Yes	Yes	Yes	Yes	No	Yes	Rechargable	US $100
Jawbone Up	Wristband	Yes	Yes	Yes	?	Yes	No	Yes	Rechargable	US $150

Table 1: Comparison table for different activity trackers

pages designed with dashboards of information. This allows the user to see the base numbers for each day as well as graphs of trends over time. Figure 3 gives an example of some of the sleep data provided by the Fitbit webpage. In addition, Fitbit provide a developer's API [14] which provides the raw data in a format which can be used within other tools or applications, and via a research partnership will also provide the data at a much lower level of granularity (e.g. steps per minute).

All of the devices have the ability to be linked to other mobile apps as well as social networking platforms to enable users to compare and 'compete' with their friends or incorporate activity tracking into other lifestyle activities such as calorie counting or fitness goals. These social aspects are not useful for our current work and we do not consider them further in this paper. In order to track sleep the devices typically need to be switched into sleep mode (apart from the Polar Loop) which requires either a button press or series of taps on the device. Both the Jawbone UP and Fitbit devices can also infer sleep if not put into sleep mode – we discuss this further later.

Focus of our studies For our long-term studies with forestry workers, we are more interested in overall trends and consistency of data recording rather than accuracy of actual step counts. Because the studies were designed and performed iteratively as the devices became available on the NZ market, not all devices and apps could be used at the same time and for all comparisons. In the following sections we describe the studies we have undertaken to consider both actual accuracy and general trend accuracy for both activity tracking and sleep monitoring and comparisons between different devices. Our aim for these studies was to determine which device(s) would be most suitable for use in initial pilot studies with forestry workers as well as to gain a better understanding of limitations and data provision.

Our short ad-hoc calibration test of trackers and apps were performed in the manner of reliability and validity tests, such as [10, 15, 16]. The 6, 10 and 12 week studies were performed as hybrid studies obtaining both qualita-tive data from diaries as well as quantitative sensor data.

6 Interpreting the Data

There is already a considerable amount of research into the effect on mental and physical performance from both physical fatigue and lack of sleep/poor sleep quality. For example work such as [17] investigates how sleepiness impairs sustained attention and vigilance as well as speed and accuracy of short-term memory and reaction time, while [18] looks at the cumulative effects of poor sleep on the functional cortex. In addition research on fatigue and tiredness in athletes, such as [19] for example, shows the effects of fatigue on both physical and mental performance. It is our intention to use the understanding already gained from such research to interpret the data we gather and begin to investigate how the regular activity levels of forestry workers over time, coupled with their sleep patterns, may contribute to a hazardous workplace environment. Additionally specific tasks, such as driving or tasks requiring hand/eye co-ordination (such as using a chainsaw) can be related to specific research in these areas. In particular, we evaluate the devices and analyse the gathered data according to the following criteria:

Data Accuracy This measure tests to what extent the data obtained reflects the activities performed. As we are planning to utilise the trackers outside of their intended use, we will be measuring both usage scenarios of (a) a typical user with sedentary office job, and (b) a forestry worker with outdoor jobs involving walking, tree felling and clearing and driving.

Data Expressiveness We are not only interested in accurately measuring steps and sleep quantity and quality, but also in identifying activities. This measure will describe the ability to identify typical activities such as driving, walking, chainsawing, tree felling and clearing based on the data obtained.

Usability Usability of devices and applications will be measured with regards to interaction design for direct and indirect interaction while measuring steps and sleep, as well as in regards to device maintenance, handling, suitability for outdoor use and motivation for regular long-term engagement.

7 Initial Investigation

Each of the devices and apps uses slightly different technologies and algorithms to measure steps and calculate calorie burn and sleep quality.

Functionality. The Fitbit devices and the Withings Pulse all use a 3-axis accelerometer sensor (the One and Pulse also contain an altimeter to monitor stair climbing), the Polar Loop and Pebble have a 3D accelerometer sensor and the Jawbone UP has a motion sensor. Each device then uses its own algorithms to turn the motion information into step calculations and includes information about the wearer (entered when the user sets up the device for the first time) such as age, height, weight etc. to calculate calorie burn. Sleep times and quality are measured based on movements and micro-movements with the devices and apps each using different measurements and algorithms to determine what particular movements mean with respect to sleep (light sleep, deep sleep, being awake etc.) as well as measuring overall sleep length from when sleep mode is started to when it is stopped (apart from the Polar Loop which determines sleep based on movement alone and does not need to be switched into sleep mode). So, for example, once in sleep mode, the Fitbit Flex counts significant movements (such as rolling over) as being awake. There are also differences in the main focus of the sleep tracking

Participant	Forestry worker	Sex	Devices	Tests	Steps	Sleep	Driving	Chainsawing	Diaries
P1	No	F	Fitbit One Withings Pulse Jawbone UP	Calibration	Yes	No	Yes	No	No
P2	No	F	Fitbit Flex Pebble Jawbone UP SHealth app	Calibration	Yes	No	Yes	Yes	No
P3	No	M	Fitbit Flex Pebble Jawbone UP SHealth app	Calibration	Yes	No	Yes	Yes	No
P4	No	M	Fitbit Flex Jawbone UP SleepAs app	12 week study	Yes	Yes	No	No	Yes
P5	No	F	Fitbit One SleepCycle app	10 week study	Yes	Yes	No	No	Yes
P6	No	F	Jawbone UP Fitbit Flex SleepAs app	8 week study	Yes	Yes	No	No	Yes
Researchers	No	F	All	ad hoc	Yes	Yes	Yes	Yes	No
FP1	Yes	M	Fitbit Flex Jawbone UP	5 week study	Yes	Yes	No	No	No
FP2	Yes	M	Jawbone UP Polar Loop	5 week study	Yes	Yes	No	No	No
RB1	No	F	Readiband Jawbone Up	1 week study	No	Yes	No	No	No
RB2	No	F	Readiband Jawbone Up	1 week study	No	Yes	No	No	No

Table 2: Overview of study participants

between devices and apps. The Polar Loop primarily focusses on length of time the user is asleep whereas the FitBit devices focus on overall sleep quality based on both length of sleep and calculations of sleep quality. The phone apps often include 'gentle alarms' which aim to wake a user gradually (through gently vibration and increasing volume alarms). The data produced by the different devices and apps and the way this is provided to the users reflects these differences.

Testing correct activities. Table 2 gives an overview of all study participants and the devices and activities that were involved in testing. We began by performing basic calibration tests involving all devices with three different participants (P1 to P3) to see how closely the activity tracking matched to actual steps taken when walking and running for very short distances. Each participant took a fixed number of steps (10, 20, 50 and 100) while walking (and in the case of P1 also running) and recorded how many steps the different trackers measured for each of these. This was to get an initial understanding of how consistent each of the devices were and whether different users had different experiences (for example the wrist-worn devices are affected by the amount of arm movement the user has when walking and are more sensitive to different 'false positives' in terms of step counting than those worn clipped to a waistband or carried in a pocket.) Most of the devices proved both accurate and consistent for these small-scale tests, both walking and running, although some anomalies were seen due to the way some participants performed the test. For example, P1 performed the calibration activities on a treadmill and found that the actions required to perform the syncing for the Jawbone UP (removing the device from the wrist, removing a cap from the end of the wristband and then plugging into a mobile phone) led to a lot of movement of the device, which in turn led to over-counting. Typically this is unlikely to be a problem for a user who is only syncing the device every 4-5 days, as any over-counting of steps will be too small to have any impact on the total data. P2 used the trackers while walking

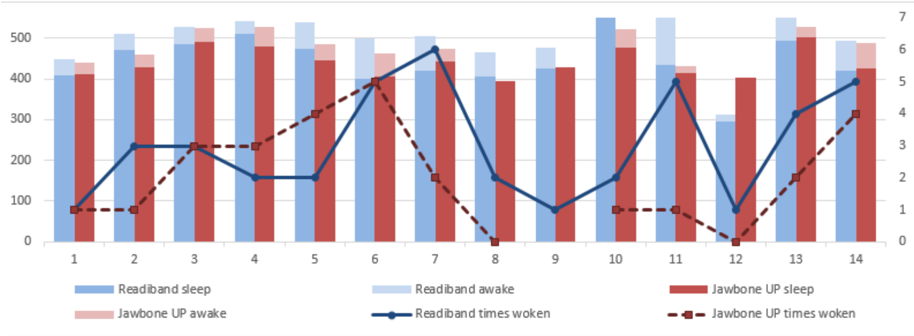

Figure 4: Readiband vs Jawbone UP (sleep time in minutes)

in a forest, and observed that particularly when stepping very carefully or walking on soft ground covered with pine needles, no steps would be counted.

Testing false activities We also needed to understand how some of the typical (non-walking) activities performed by forestry workers, such as driving, using a chainsaw etc., may affect the step counting on trackers, and so conducted a number of smaller 'ad-hoc' tests to measure these, carried out by the researchers. During the calibration tests, P3 had already observed that the Pebble smart watch counts large numbers of steps when a wearer was driving a car, so it was important to see whether this was a general problem. Also both arm movements and vibration from chainsaw activities were likely to register movement and we similarly needed to see what sort of levels were recorded with these activities. While using a chainsaw to cut up tree stumps and larger branches, we tested for potential false step-counting on the Jawbone UP, Fitbit Flex, Pebble and S-Health app. We observed that the vibrations of the chainsaw did not trigger steps to be counted (i.e., no false positives) on the Fitbit or Jawbone. However, jerky arm movements did trigger the step counters on the Pebble and the S-Health app.

We also tested how long drives (typical in the forestry industry) influence the tracked activities. We observed that S-Health did (correctly) not count any steps on a drive of more than 100km (taking approximately 75 minutes). However, all wrist-worn devices did falsely count steps while driving (Fitbit Flex counted about 500 steps on a 100km drive, the Pebble reported about 1000 steps for the same distance and the Jawbone UP counted 351 (driving tests used automatic transmission vehicles). The Fitbit One and Withings Pulse (both in right hand trouser pocket) counted 300 and 305 steps respectively. Closer examination revealed that it was not, as suspected, arm movements while driving that set off the step counter, but rather, uneven tracks and any bumps or holes in the road triggered most of the false positives.

Our final calibration tests involved a comparison on the sleep monitoring of the Jawbone UP with the Readiband. Previous studies have shown that sleep measurements from activity trackers typically over-estimate length of sleep and are poor at detecting sleep anomalies [6] compared to polysomnography. The Readiband is described as the "most accurate way to measure and quantify sleep outside of a clinical sleep lab" [20] and uses algorithms developed by the US military. The Readiband works with a proprietary software and only measures sleep (not activity), both its cost and functionality mean it is not suitable for use within our studies, however we were interested in comparing the data it produced with that of an activity tracker. Two participants wore both a Readiband and Jawbone UP on the same arm for a period of one week (see Figure 4, P1 for day 1–7, P2 for day 8–14). The data from day 9 is a sleep estimate by the Jawbone (due to low battery). We observe that the Jawbone UP consistently underestimates the sleep time (bar once), but is overall detecting similar patters to the Readiband. The Jawbone UP's detection of awake times was found to be less reliable.

Discussion. Most of the devices provided consistent activity tracking for the short calibration tests and, apart from the Pebble and S-Health app, were not affected by the chainsawing activity. Driving may generate false activity data, which is particularly relevant as forestry workers often drive long distances both as part of their working day and to travel to and from work places and much of these distances are over bumpy secondary roads or tracks. Being able to include driving data will be useful as it contributes to the workload of the workers and requires at the very least good cognitive response times. Being able to identify it separately from other activities as "driving" may also prove useful, however, at present it is unclear whether a metric such as 500 steps per 100km as recorded, is in any sense an accurate estimate of the effect of such driving with respect to fatigue. This will require further study to determine. Using the activity trackers alone, we cannot determine which activities are causing the trackers to record movement – so 20,000 recorded steps may be from any combination of walking, driving and chain-sawing or other work-related activities. At present we are not interested in separating these different activities out so that they can be measured independently we are, rather, interested in over-all activity levels of the workers throughout the day. As such combining all of these values is not problematic as they allow us to firstly gain an understanding of what typical activity levels are and secondly establish a baseline of worker activities which can then be compared over different days.

8 Three Longer Studies

The goal of the longer studies was to consider how consistent general trends were across devices. For example even if a device does not count steps with 100% accuracy for a particular user, is it the case that it remains consistent so that the patterns between different users and different devices remain similar over longer periods of time? Previous comparison studies of individual devices, such as [21, 22] for example, have looked at issues such as data comparison and usability aspects (battery life, synching, supporting apps etc) but typically over much shorter periods of time and with the aim of identifying which device is better for a particular type of consumer. We performed three longer studies with three different participants (P4, P5 and P6) in order to compare aspects of some of the trackers and phone sleep apps.

Our first study consisted of a 12 week comparison between the Fitbit Flex and Polar Loop with respect to steps counted, calories burned, sleep tracking and battery life and also included a comparison with the Android Phone app SleepAs. P4 wore the Fitbit Flex and Polar Loop on different arms, 24 hours a day for the 12 week period and also used the phone app every night for the same period. We had three aims with this study. The first was to see if the two activity trackers were consistent with each other for each of the three data items and how these compared to the sleep data of the phone app. The second was to identify whether there were differences depending on which arm the tracker was worn on (dominant or non-dominant [3]). The participant swapped these over half way through the study. The third was to identify any usability issues that occurred with either device when worn 24 hours a day for a 12 week period which

[3] the Fitbit Flex allows the user to set this parameter but the Polar Loop assumes it is worn on non-dominant arm

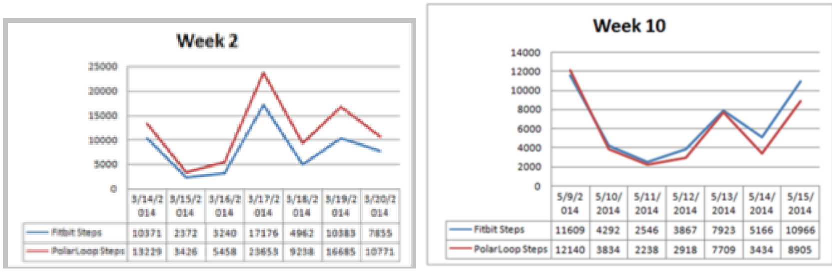

Figure 5: Steps Counted – Fitbit Flex vs. Polar Loop

might affect their suitability for our larger scale studies. The second study was a 10 week comparison between the sleep data gathered by the Fitbit One and an iPhone sleep monitoring application called Sleep Cycle. P5 carried the Fitbit One every day for the 10 week period and used it in the sleep armband every night as well as using the Sleep Cycle app every night for the same period. The third study was an 8 week comparison of activity and sleep monitoring between the Jawbone UP, Fitbit Flex and SleepAs Android phone app. P6 wore the Jawbone UP and Fitbit Flex on different arms for the 8 week period and used the Sleep As app every night. This study also included an exploration of the use of a companion phone app S-Health for step counting.

Participants in all three studies kept diaries where they recorded data such as their personal perception of how well they had slept the previous night and any contributing factors that might affect activity levels (such as playing sport or spending all day in meetings). In addition they recorded any usability issues or problems that occurred and kept track of how often they charged each device.

Data Results All three studies counted steps with various trackers. Figure 5 shows example graphs from the data gathered during weeks 2 and 10 by P4. While there are differences between the steps counted by the Fitbit Flex and Polar Loop, the difference between them remained reasonably consistent. During the second half of the study (weeks 7-12) the user swapped the tracker to the opposite arm, the difference between steps counted was much smaller. The largest margin between steps counted by P4 was in week 4, when the Polar Loop recorded 8,436 more steps than the Fitbit Flex; in addition the Polar Loop was on charge for one hour of this day so the actual discrepancy was larger. The two main activities recorded in the user diary were attending a party and going out drinking with friends. It was inferred that the extra steps recorded were actually likely to have been caused by increased arm movements on the dominant arm (which the Polar Loop was being worn on) relating to drinking. The participant also observed that the day he took the highest number of steps throughout the study (and received congratulatory messages from the Fitbit Flex tracker) was due to being on a pub crawl. Collecting context-free data means that there may be reasons for high step/activity counts which are not directly related to number of steps taken by a participant, or are not related to 'fitness' type activities.

The sleep data contained the biggest differences both between different activity trackers (worn on the wrist while sleeping) and the mobile phone apps (reliant on the phone being placed next to a pillow). The devices and apps track such metrics as total time asleep, number of times woken and total time awake during the night and then apply a quality score to the overall data. For P5 the Jawbone UP measured the number of times the participant woke during the night at an average of 3-4 times per night over

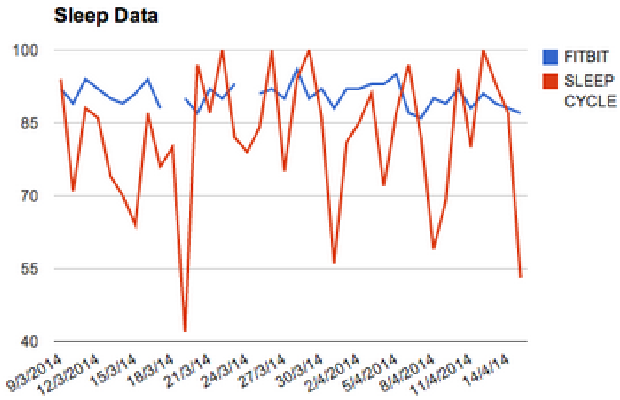

Figure 6: Sleep Quality comparison Fitbit One and Sleep Cycle

a ten week period, while the FitBit One for the same participant showed an average of 20 times. Comparing the data with the participant's diary record suggests that the FitBit One is much more likely to record light sleep or motion as being awake, consistently in contrast to the Jawbone UP measurements which more accurately reflected P5's own rating of her sleep. Quality scores of sleep for the Fitbit One were, however, consistently higher than those given by the Sleep Cycle app. In addition, sleep data was more prone to be missed from the logs for a variety of reasons, such as the participant forgetting to put the device into sleep mode correctly, the participant lying on the arm on which the device was being worn which caused the button to be depressed which turned sleep mode off, a lack of battery power prior to going to sleep, and a device failing to recognise it had been put back on after charging.

Figure 6 shows some of the comparison data between the Fitbit One and the Sleep-Cycle mobile app from P4, which highlights these differences. Note the gaps in the graph, which indicate missing data. Figure 7 shows sleep data for two nights from both Jawbone UP and SleepAs. The bright green phase (bottom graph of Figure 7) on SleepAs indicates that P6 paused the sleep tracking while she was awake. In comparison it can be seen that the Jawbone correctly detects awake times (denoted by the orange blocks) while the SleepAs app only detects light sleep. Comparing Fitbit, Jawbone and SleepAs to the subjective experience of P6, the Jawbone detected sleep cycles and awake times most accurately (or at least best matching the subjective experience of P6).

For one night, P6 had the opportunity to compare sleep tracking using two SleepAs mobile apps on different mobile phones, one of which used the in-built sensors and the other one used data provided by a wrist-worn Pebble smart watch (see Figure 8). The second phone that was used with the app alone had not been used before. It seemed much less responsive and therefore detected fewer light sleep phases than usual. The phone that was connected to the Pebble detected significantly more movements than usual, thus reporting many light-sleep phases. Further research is needed to explore the use of smart watches in combination with phone apps.

Data Problems Identified All of the devices use aspects of motion to monitor sleep and determine level of sleep (e.g. little or no movement indicates the user is in a deep sleep whereas lots of movement indicates either light sleep or wakefulness). P5 noted

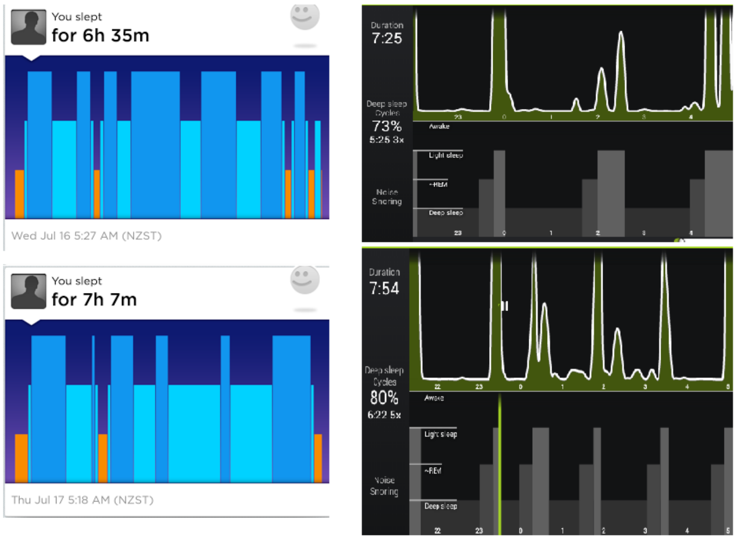

Figure 7: Jawbone vs SleepAs

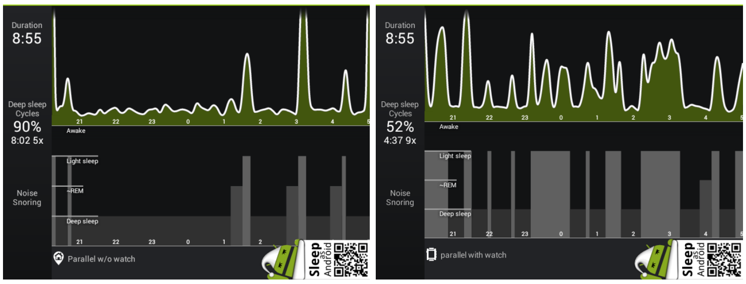

Figure 8: SleepAs using phone only (l) and with Pebble (r)

that if she was awake early in the morning (about an hour before her alarm went off) she would typically stay in bed and use this as ‘thinking time’. She would be almost motionless during this period and the Fitbit One would therefore record this as deep sleep rather than the total opposite of being awake which was actually the case. Another surprising result was that if P6 interacted with the mobile sleep application (opening up the application to view aspects of the recording) it did not recognise that this meant she was awake, and would continue to record sleep until she actually activated the ‘end sleep’ mode. The sleep data was much less consistent than the step counts and often the quality scores and wakefulness metrics did not reflect the participants’ diary observations of how well they thought they had slept or how rested they felt.

Both Fitbit devices and Jawbone UP provide features which enable a user to track sleep even if they have forgotten to put the device into sleep mode. The Flex and One allow the user to enter start and end times for sleep and then retrospectively calculate the sleep quality metrics based on activity data gathered during that period. The Jawbone UP works in a similar fashion but suggests the start and end times based on lack of activity and allows the user to either accept the suggested time or manually enter their own times. We performed an ad-hoc test on this feature where one of the researchers wore 2 devices on the same arm at night, putting one into sleep mode but not the other. We then compared the data calculated from sleep mode against the data inferred from start and end times of sleep. The data produced by the two devices was almost identical, which was promising. However, we have encountered numerous examples on both the Fitbit Flex and the Jawbone UP where this feature does not work, or is not enabled, without any clear or obvious reason why, suggesting it cannot be relied upon.

Usability issues and problems uncovered There were some minor issues encountered with wearing the devices during particular activities. P4 found that the size of the Polar Loop meant that if he was doing computer work and typing on the keyboard then the wristband would press into his arm causing discomfort. The same participant also found that the Jawbone UP had a tendency to get caught on clothing when getting changed (the UP does not have a closed wristband but rather open ends which cross over each other to allow for different size of wearers' wrists). P5 found that if the Fitbit One sleep band was not carefully fitted around her wrist at night the velcro fastener rubbed and caused a rash. P5 also found that on more than one occasion she forgot to remove the Fitbit One from a piece of clothing prior to it being washed; luckily the device is waterproof enough to survive this.

Another issue identified was that of motivation, particularly around sleep monitoring. Of the devices being tested, both the Fitbit Flex and One work best if they are manually put into sleep mode (the Polar Loop infers this from movement) and the One needs to be removed from its clip-on cover and put into a soft band that the user then wears around their wrist. As our participants were also testing mobile sleep apps they also had to start the sleep monitoring apps and put the phones in the correct position (either under the sheet or under the pillow) before going to sleep. At around the 6-7 week mark all participants found that at times these procedures were annoying, particularly if they just wanted to go to bed after a long day or late night. Although our forestry workers will only have one device to consider (and therefore this should be less of a problem) they may be less motivated to 'remember' to perform the relevant actions needed prior to going to sleep. This means that even minimal overhead to perform the task of putting a device into sleep mode cannot be ignored. The requirement to connect the mobile phone to a power outlet during sleep tracking is only possible if the user has a suitably placed power supply and the length of the charging cord is long enough to enable correct positioning of the phone once connected. As the applications tested both work irrespective of whether or not the phone is plugged in (it is a recommendation rather than a hard requirement) this is not an insurmountable problem, but users must ensure enough battery life before going to bed if they are unable to connect to a power supply. This means that wearers need to keep track of battery levels (in activity trackers as well) to ensure they charge at appropriate times (rather than just waiting for the battery to go flat), which again adds to the overhead for participants in long-term studies.

9 Pilot Studies

The next step was to proceed with evaluating the activity trackers with forestry workers in their working and home environments. The full set of pilot studies will involve different teams of forestry workers (which vary in size from 2 to 10) wearing trackers for periods of time ranging from two to six weeks. Our first study, however, was intended as a way of introducing ourselves to one of the contractors and his team and gaining some understanding of how to set up and run studies in the forestry environment as well as to discover any immediate or initial problems with workers using the devices. As such this first study involved only two workers.

The two forestry workers (FP1 and FP2) who are involved in 'thinning' (using chainsaws to remove trees from densely planted areas) were supplied with activity trackers to wear for five weeks. Based on our initial device studies we decided to start with two of the wrist worn trackers – FP1 was given the Fitbit Flex and FP2 was given the Jawbone UP. We chose wrist-worn trackers over those that are carried or clipped onto clothing as we believed it would help ensure the workers would wear them for 24 hours a day. We had already seen in our early studies that it was easy to forget about these types of trackers and leave them in pockets so it was equally likely that works would forget to transfer trackers from work clothes onto casual clothes and vice versa. Also, the overhead of putting the devices into sleep mode was lower (these trackers did not require placement into a separate wrist band). The two participants were briefed at the start of the study to explain the purpose of the study and anticipated activities over the ensuing weeks. The two participants had slightly different roles: FP1 was responsible for the majority of the driving, both were involved in clearing areas of scrub around trees to be felled, and FP2 was responsible for the majority of the tree felling. Once thinners have felled trees they leave them on the ground (it is not their responsibility to remove them). The workers estimate that they clear around 500 trees a day (with average tree size being 20+ m high with a trunk circumference of 18+ inches).

At our first meeting with the participants we gathered some demographic data (age, length of time as a forestry worker, major working activities) and explained how the different activity trackers worked. Both participants have worked in the forestry industry for around 15 years and are in the 30-35 years age bracket. They both own smartphones (FP1 has a Samsung smartphone and FP2 an iPhone). The participants were also provided with a one page information sheet which explained how to put the device into sleep mode and how and when to charge the device. One point of interest during this initial meeting was the immediate concern of FP1 about 'looking after' the activity tracker. He was worried that it might get dragged off his wrist and lost or damaged while walking through the bush and suggested putting it in his pocket instead. It was important to reassure both participants that if that happened it would be useful data in itself (i.e., we would learn that a particular device was not suitable for use in that environment) and that would be more useful to us than if they took them off to protect them or put them in their pocket, as that would affect data collection and increase the likelihood that they would forget to wear it outside of working hours. This reinforced our initial belief that using low-cost devices which could be considered 'disposable' is essential for successful data collection.

Our plan was to contact the workers by phone every couple of days to check in with them and make sure everything was running smoothly. During our initial meeting we had identified suitable times to contact them and also the best way to make contact. Despite this we were not always able to make contact with them when planned and only managed to speak to them once during this first week. It subsequently transpired that sending them a text was more likely to elicit a quick response (usually along the lines of "Have you managed to successfully charge the tracker this week?" with an immediate response of "Yes"). At the end of the first week we met with the participants to gather the initial set of data and discuss how things were working for them so far. We then returned at the end of week two; we had planned to end this initial study at that point, but based on early findings decided to extend it for a further three weeks. We now discuss the data and findings.

10 Initial Results from Pilot

FP1 lost his tracker on day 3, as it had come off his wrist while working, he had, however, found it again the next day and put it back on. Neither participant had charged their device (despite the info sheets requesting charging on days 5 and 7 respectively) and one of the devices was almost out of battery power. FP1 had failed to ever put his device into sleep mode (Fitbit Flex) and was confused by the 5 light display which he thought meant the device was always in sleep mode (actually it never was). FP2 had managed to successfully track 3 out of 7 nights of sleep and the UP had also been put in and out of sleep mode at random times during the day and night several times. We were able to use the retrospective sleep tracking functions of both the Flex and UP to fill in some of the missing sleep data gaps but not all of it; in addition the partial recording of sleep by the UP at random points during the day affected some of FP2's results. The activity data gathered showed that FP1 had an average daily step count (during working week) of 13,688 steps (approx. 9.6km while FP2 had an average daily step count (during working week) of 22,736 steps (approx. 16km). We were able to clearly identify phases of low physical activity (e.g., while driving) and walking. The sleep data gathered was too inconsistent to use for further analysis. Some other problems we encountered were that after the battery of the Fitbit Flex had run down completely and been left for over 24 hours to be recharged the internal time zone reset to US time (during initial setup it had been set to NZ time) which meant that most activities for the following week were incorrectly timed as occurring during the night.

It was clear from these first observations that devices with the least possible participant interaction will be required if we are to successfully collect large amounts of data from large numbers of participants. In addition to low-cost and disposability, our main requirements are now that little or no interaction is required to track sleep and the longest possible battery life is required. As a result of the sleep data inconsistency we decided at the end of week two to replace FP2's Jawbone UP with a Polar Loop (which automatically infers sleep mode) to see if that would improve data collection, and to swap FP1's Fitbit Flex with another Jawbone UP which is easier to put in and out of sleep mode. During the second part of the pilot in weeks 3-6 the Jawbone UP recorded no data for the first ten days, and then afterwords tracked all activity and half of the sleep data. FP1 again lost his tracker whilst scrub clearing, but once again was able to find it soon after. When we collected the devices at the end of week six FP2 forgot the bring the connector for the Polar Loop (which enables charging and data syncing) and at the time of writing we are waiting for this to be returned so that we can collect the data from the device.

11 Discussion

Similar to our previous studies, we found that sleep tracking was the most unreliable in terms of the ability to record this data. Wrist-worn activity trackers are prone to getting caught on branches during scrub clearing activities which can cause them to come off, and so for larger-scale studies we will need to build in redundancy (in terms of both participant numbers and replacement devices) to compensate for this possibility. Anything which requires user interaction (putting the device into sleep mode, charging the device etc.) is also problematic. We had considered asking future forestry pilot participants to try synching the devices to mobile apps themselves (we performed this task ourselves at the end of weeks 1, 2 and 6 when we met with FP1 and FP2) to upload the data. However, based on this first pilot study it seems that this may prove problematic, particularly for the Jawbone UP as it must be physically connected to a phone for the sync to take place (the Flex can sync via Bluetooth when the tracker is in close proximity to a phone with the Fitbit app installed). These initial findings also suggest selecting devices with the minimum amount of interaction required and longest possible battery life. It is common in studies such as ours to include 'self-reporting' aspects (as we did via the use of diaries in our initial device evaluations), however we believe these may be problematic with forestry workers. We have already seen that contacting workers regularly (even when they choose the best way and times for that contact to take place) is not easy. As such, finding appropriate ways to for the workers to provide useful information (which may include how they feel wrt. tiredness, fatigue etc.) in an easy and convenient way that they will remember to do is another area that requires further investigation.

We must also remain mindful of the ethics of collecting the sorts of data we are interested in and how (and by whom) it is used. While our goal is to try and contribute to a safer working environment for forestry workers, the use of data to identify workers who may not be in a suitable physical or mental state to work may lead to a reluctance on the part of the workers to participate for fear they may be stood down or worse, lose their jobs. A case study presented in [1] describes how one company uses spot-checks and issues yellow (warning) and red (stand down) cards to workers who break safety rules. It is suggested that this is supported by the workers as it is part of a larger safety emphasis (which also includes bonuses and rewards to workers based on attendance, productivity and lack of safety breaches). It is important that our work receives similar support from workers and is used in a way which can equally been seen to benefit them rather than as a punitive measure.

12 Conclusions and Future Work

This paper described research undertaken as the first step in a larger project of using lightweight, wearable activity trackers to gather large-scale data within the forestry industry. The research enabled us to identify several important pieces of information which will inform the next step of our work, notably:

- Accuracy and consistency of step tracking is acceptable across all tested devices
- Sleep tracking is problematic for most devices, both in terms of accuracy of data gathered and the likelihood of data not being collected at all during longer studies
- Chainsaw activities do not significantly impact step counting in most trackers but driving can affect wrist-worn trackers, and activities involving significant arm motion (drinking, eating) can also make the counts inaccurate
- The use of cheap devices which can be considered 'disposable' is crucial in the forestry environment due to both the likelihood of loss or damage as well as to avoid workers taking steps to 'protect' them (such as removing them and putting them somewhere safe).
- Devices which can be worn and forgotten (i.e. requiring minimal interaction) are better suited for forestry workers

It is clear that more work is required in the area of sleep tracking. One aspect we are currently investigating is looking at other ways of determining the actual affects of sleep patterns on physical and cognitive response times. We are currently working with sports scientists to develop a response-monitoring app which will be used for this. Despite the problems we have identified, these initial studies suggest that we can use light-weight, low cost activity trackers to collect data in the manner we have proposed. While there are still further investigations required to ensure this will be successful, we are confident that the work will be possible and that the data will prove beneficial to the long-term goal of improving safety for forestry workers.

References

[1] Adams, G., Armstrong, H., Cosman, M.: Independent forestry safety review. Final Report (2014)

[2] NZ Herald: New forestry fatality takes 2013 death toll to 10 (2013) 19 December 2013.

[3] Chief Coroner's office: Recommendations recap, volume 4, July-September 2012, online report

[4] Sunday Star Times: Sports invention to save workers' lives (2014) 16 Feb 2014.

[5] Parker, R.: Technological advances in the analysis of work in dangerous environments: Tree felling and rural fire fighting. PhD thesis, Massey University, Palmerston North (2010)

[6] Montgomery-Downs, H., Insana, P., J, B.: Movement towards a novel activity monitoring device. Sleep and Breathing **16** (2012) 913–917

[7] Freedson, P., Bowles, H., Troiano, R., Haskell, W.: Assessment of physical activity using wearable monitors. Medicine and science in sports and exercise **1** (2012)

[8] Matthews, C., Hagstromer, M., Pober, D., Bowles, H.: Best practices for using physical activity monitors in population-based research. Medicine and science in sports and exercise **1** (2012)

[9] Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. **46**(3) (January 2014) 33:1–33:33

[10] Lee, J.M., Kim, Y., Welk, G.J.: Validity of consumer-based physical activity monitors. Medicine and science in sports and exercise (2014)

[11] Sasaki, J.E., Hickey, A., Mavilia, M., Tedesco, J., John, D., Kozey Keadle, S., Freedson, P.S.: Validation of the fitbit wireless activity tracker® for prediction of energy expenditure. J Phys Act Health (2014)

[12] Boon, R.M., Hamlin, M.J., Steel, G.D., Ross, J.J.: Validation of the new zealand physical activity questionnaire (nzpaq-lf) and the international physical activity questionnaire (ipaq-lf) with accelerometry. British journal of sports medicine (2008) bjsports52167

[13] Downs, A., Van Hoomissen, J., Lafrenz, A., Julka, D.L.: Accelerometer-measured versus self-reported physical activity in college students: Implications for research and practice. Journal of American College Health **62**(3) (2014) 204–212

[14] Dev Fitbit: Fitbit Developer's API: http://dev.fitbit.com/

[15] Matthew, C.E.: Calibration of accelerometer output for adults. Medicine and science in sports and exercise **37**(11 Suppl) (2005) S512–22

[16] Macfarlane, D.J., Lee, C.C., Ho, E.Y., Chan, K.L., Chan, D.: Convergent validity of six methods to assess physical activity in daily life. Journal of Applied Physiology **101**(5) (2006) 1328–1334

[17] Alhoa, P., Polo-Kantola, P.: Sleep deprivation: Impact on cognitive performance. Neuropsychiatric Disease & Treatment **5** (2007)

[18] Durmer, J., Dinges, D.: Neurocognitive consequences of sleep deprivation. SEMINARS IN NEUROLOGY **25**(1) (2005)

[19] Reilly, T., Edwards, B.: Altered sleep-wake cycles and physical performance in athletes. Physiology and Behavior **90** (2007) 274–284

[20] Fatigue Science: Fatigue Science Readiband : http://fatiguescience.com/solutions/readiband/

[21] Ari Jay: arijaycomet.com Activity tracker comparisons: http://arijaycomet.com/category/health-fitness/

[22] Guo, F., Li, Y., Kankanhalli, M.S., Brown, M.S.: An evaluation of wearable activity monitoring devices. In: WS on Personal Data Meets Distributed Multimedia, ACM (2013) 31–34