Intersect, Vol 6, No 2 (2013)

Phases of Accuracy Diagnosis: (In)visibility of System Status in the Fitbit

Molly Mackinlay
Stanford University

This paper analyzes the Fitbit—an all-in-one step, floor, distance, calorie, and sleep tracker—to explore user perception of accuracy in black-box systems (systems in which the user has no insight into the device's inner workings). Due to learned expectations of system unreliability, users are skeptical of Fitbit recorded data. Because of this, many users perform tests to understand the Fitbit's level of precision, ultimately revising their mental model of the Fitbit itself and attempting to calibrate their personal use of the device. However, due to the limited visibility of the system's functioning status, efforts to test and calibrate Fitbit data are ultimately flawed. This paper examines the results of seven interviews with Fitbit users, theorizes and describes four phases of use many Fitbit users undergo, and concludes with a critique of the usability of the Fitbit.

Technology

The Fitbit is a thumb-sized activity tracker intended to be worn 24 hours a day. Users clip it to the waistband of their pants during the day to count their steps and wear it on a Velcro wristband at night to track their sleep. Users can view their recorded data in three ways: through an iPhone application, a web application, or through the Fitbit itself. The iPhone application—to which the Fitbit connects wirelessly—simply displays the recorded numbers (steps taken, calories burned, floors climbed, etc), while the web application presents the data in slightly greater detail and visualizes trends in data over time. Pressing a small black button on the Fitbit causes the display to light up with the number of steps taken; pressing this button repeatedly will cycle through the values being recorded, each distinguished by a descriptive icon in the upper right corner. If this button is pressed and held, a count-up timer will appear on the screen. This denotes the time spent in sleep tracking mode and can be turned off by once again pressing and holding the button.

FIGURE 1. Fitbit One Wireless Activity and Sleep Tracker.

Importance

Increased interest in health and activity tracking has spurred the adoption of wearable fitness technologies such as the Fitbit, the Nike FuelBand, and the Jawbone Up wristband (Olof, 2013). At a time when 35.7% of the US population qualifies as obese, these technologies are helping individuals concerned with living healthier lifestyles quantify and track their daily activity and set fitness goals (Center for Disease Control and Prevention, 2012).

From the interviews in this study and related literature in the media, it is clear that the Fitbit's data recording truly does encourage some users to exercise more (Guzman, March 2013; Guzman, January 2013). One user in this study commented that at around 10pm on a day when the Fitbit was reporting particularly low numbers, she chose to take a walk to increase her daily step count. On another occasion, when 100 steps away from her goal at 11:50pm, she jogged in place until her daily step goal was met. Some users highlighted examples of times their Fitbit step counts prompted them to go to the gym; others mentioned the device's mere presence motivating them to take a longer route to class, or to walk instead of bike to a meeting. These anecdotes demonstrate how the Fitbit helps foster healthier lifestyles.

Literature

Current research indicates that for wearable technology to be effective, its data must be relevant, pervasive, and understandable. Relevance for devices such as the Fitbit means the data collected must be contextualized so that users can accurately assess their progress (Richmond, 2013). In order for these devices to motivate change or inform action, they must first

provide users with relevant information about their current activity level and help them set realistic goals (Guzman, March 2013; Guzman, January 2013). The data should also be pervasive: it should be synced throughout a collection of technological devices like smart phones and computers for easy viewing and analysis (Richmond, 2013). Finally, it is important that the activity data be understandable: it should be broken down into coherent units from which users can derive useful insights. The majority of analysis about trackers like the Fitbit focuses on these three factors. However, little research has focused on the accuracy of the trackers themselves.

Methods

Interviews were conducted with seven Fitbit users. The seven users were Stanford students between the ages of 21 and 23; three interviewees were male and four were female. Two users performed think-aloud protocols (verbalizing their thought progression) while experimenting with the Fitbit for the first time. Three users were given the Fitbit for three days and asked to use the device as if it were their own, tracking their experiences and learning process. The two remaining users were experts, having logged months or years of Fitbit use.

Users performed a variety of activities over the period of this study, including walking to class, biking, running, attending workout classes, exercising at the gym, dancing, and sleeping. They were told to go about their daily lives (six of the seven live on Stanford's campus) while wearing the Fitbit. Afterwards, these users were interviewed about their impressions of the Fitbit. Questions focused on changes in use over time and moments in which users had insights or perspective changes about the Fitbit.

Interviews were conducted without any pre-determined research goals. My initial hypothesis was that users would have interesting insights about behavior change due to the Fitbit's output. Instead, the interviews revealed that users were most concerned with system accuracy and preoccupied with questions about system status and truthfulness.

Phases of Fitbit Use

Based on my interviews, the most persistent question among users was that of system accuracy. Users were unsure what constituted a *step, floor, calorie,* or *mile* to the Fitbit because they were given no ability to calibrate the device or diagnose its margin of error. To judge system accuracy, users were found to have a variety of mental models of the Fitbit's internal workings. This paper hypothesizes that these models were developed through four main phases of device use: Introduction, Trial, Personal Calibration, and Satisficing.

Phase Zero – Initial Preconceptions

Users initially approached the Fitbit with significant experience with mobile applications, including location and activity trackers, and these experiences contributed to a set of expectations about the reliability, precision, and accuracy of the Fitbit. Based on experiences with related applications like GoogleMaps, Nike+, and simple pedometers, users initially viewed wearable trackers as buggy, inaccurate, and highly error-prone. This conception of technology as inherently unreliable might be partially due to the early stages at which many of these users adopted the new technologies. Trying new and experimental applications before they are widely accepted and tested trains early adopters to be cautious and skeptical of the data reported by an application. Therefore many of our users were initially suspicious not only of the numbers reported by the Fitbit, but also of the functional definitions of the terms used by the device.

Phase One – Introduction

The introduction phase was characterized by a superficial exploration of the device's capabilities, including how to turn it on, view output, and interpret its messages. Users began building their mental model of the Fitbit while testing system reactivity.

When first introduced to the Fitbit, users often struggled. One user's first question was: "is it on?" after which she shook the device around her head while checking the tiny screen for any response. Other users cycled through the different display options by pressing the home button to get a sense of the different options the Fitbit provided. One of the users had trouble transitioning the Fitbit from the off/charging state to the tracking/responsive state. She explained that she had hardly used the device since getting it because of her inability to figure out if it was actually turned on.

These behaviors highlight two main usability problems, both of which were reinforced throughout the next three phases. The first problem, borrowing Jakob Nielsen's term from *10 Usability Heuristics for User Interface Design*, is the *visibility of system status*, meaning users were unable to diagnose the internal state of the device and predict its behavior (Nielsen, 1995). The second usability problem was the system's violation of the *Maxim of Quality* introduced in Paul Grice's *Cooperative Principle*, which describes the qualities necessary for positive interpersonal interactions (Grice, 1975). Violating this maxim caused users to mistrust the system's accuracy. These two issues caused users to become increasingly mistrustful of the Fitbit's accuracy in this phase.

The concept underlying *visibility of system status* is to always give users prompt and clear feedback so that they can diagnose what state their device is in and whether an action produced any results. With the Fitbit, users were unable to diagnose if their device was in "off" or "tracking" mode since the two are visually identical. A lack of response from the device when it was moved or a button was pressed could indicate that the

device was off or that the movement/press hadn't been of sufficient strength to activate the display. A third possibility was that the device was broken and no amount of charging would restore system functionality.

Without sufficient insight into the system's internal workings, new users were often confused about how to approach the Fitbit and elicit the expected system response. Some users were convinced their device was faulty, while others kept accidentally putting it in "sleep" mode because their presses were too forceful.

The second usability guideline violated by the Fitbit was Grice's *Maxim of Quality*. This maxim is founded upon the *Be Truthful* mandate: to "not say what you believe to be false," and "not say that for which you lack adequate evidence" (Grice, 1975, p. 46). In the introduction stage, some users waved the Fitbit in the air and watched the number of counted steps increase in order to diagnose that the device was on and working. However, this immediately cast doubt on the authenticity of the step count reported by the device since the user was able to increase the count without taking any actual steps. The user therefore concluded that the Fitbit was either misrepresenting the data collected or imprecise enough in its measurements to incorrectly record random shaking as steps. The Fitbit's violation of this maxim sabotaged its relationship with the user and undermined the user's already weak trust of the system's accuracy.

These failures caused the user to be suspicious about how the Fitbit measures a step, whether all steps are faithfully recorded, and if movements that aren't steps are interpreted correctly. Unfortunately, the Fitbit cannot answer these questions due to its minimalist design. User doubts are even stronger regarding concepts such as *floors*, which do not have consistent definitions. At the end of the introduction phase, most of the users in my study had confirmed their original skepticism of the Fitbit's accuracy and learned to regard the data provided by the Fitbit as highly suspect.

Phase Two – Trial

The trial phase consists of both explicit and implicit testing. By this point, users have explored some of the Fitbit's possible states and are attempting to use the device in their daily lives to collect meaningful insights. The results of their testing reinforce the issues encountered in phase one.

In the trial phase, users engaged in two main types of testing. Some users attached the Fitbit, counted their steps as they walked across the room, and then checked the Fitbit to see if the two numbers were in agreement. Other users simply started using the Fitbit and checked it before and after various activities to intuitively "feel" its accuracy.

Throughout this testing, users had to work around the Fitbit's black-box-like nature and lack of testing features. It was hard for users to use the device as prescribed in the instructions while also keeping the screen that was reporting step or floor count in view. Users also bemoaned the manual nature of testing as they tried to remember arbitrary step/floor numbers

before and after different activities, since the Fitbit has no "reset" or "clear" function.

Phase Three – Personal Calibration

The personal calibration phase consisted of goal setting and subtle error discovery. Successful users learned to use Fitbit data not as absolute, but as relative measurements. Their continued cautious use also led many users to discover subtle errors or misdetections by the Fitbit in certain activities or contexts. This progression of use is consistent with a so-called "low floor, low ceiling" system, in which a device is easy to begin using but has minimal space for improvement. This style of learning curve is said to be "shallow" since there is little difference between expert and novice use. (Resnick et al., 2013, p. 3)

Most users set goals by either tracking daily numbers for a few days before setting an appropriately challenging but achievable goal, or by iteratively setting higher and higher goals until the goal is sufficiently challenging. These methods were most popular specifically due to the doubt created by the Fitbit's violation of system status visibility and the Maxim of Quality. In this phase, it didn't matter if the Fitbit gave the actual "right" number as long as it gave relative numbers to track progress toward relative goals.

Interviews revealed that after wearing the Fitbit for a few days, users discovered subtle activities or contexts in which the device was particularly inaccurate. One user noticed the "distance gone" numbers were always far off her expectations when she went running on the treadmill and that floors climbed didn't count on StairMasters. Another user noticed the Fitbit was nearly useless at tracking biking around campus or in the gym. Conversely, one user found the Fitbit heavily over-counted steps while the user performed Zumba, possibly due to the many hip movements characteristic of Latin dance. These incremental discoveries allowed users to work around anomalous data, but users were ultimately unable to modify the Fitbit to account for these inaccuracies in the future.

The users' inability to calibrate their device or improve its accuracy resulted in the device having a low ceiling for use. An "expert" Fitbit user is nearly identical to a "novice", since the system offers no explanation as to *why* an activity was poorly measured, and there exist no configuration tools to improve future tracking. While having a low floor is important for market penetration, the Fitbit's individual impact and long-term retention are severely limited by its low ceiling and shallow learning curve.

Phase Four – Satisficing

After using the Fitbit for a few weeks, most users progress into the final stage of use: satisficing. At this point users know generally what activities are well tracked, have lost their initial infatuation with the device, and have serious doubts about the Fitbit's accuracy.

Colloquial interviews showed that users learned where on their bodies

to wear the device for most accurate tracking, along with what activities weren't worth tracking for them. Many users stopped wearing the device regularly, while others simply stopped monitoring their data as obsessively.

In other products, expert users often experience an (approximately) exponential learning curve. Due to the Fitbit's low ceiling, Fitbit users, however, do not. The Fitbit offers neither advice on how to improve accuracy nor methods to decrease its margin of error. While some expert users might develop workarounds for the Fitbit's flaws, this fundamental inability to improve data quality handicaps the Fitbit's success.

Conclusion

This paper has demonstrated how the Fitbit's effectiveness is fundamentally limited by its violation of the *Maxim of Quality*, its lack of Visibility of System Status, and its low-ceiling learning curve. While the Fitbit may still encourage positive fitness habits, users are inclined to abandon their use of the Fitbit because: they don't trust its accuracy; they have trouble understanding what state it is in; and they are unable to make the Fitbit more accurate as they become expert users.

Understanding why activity trackers are abandoned is important not only to health technology designers but to the entire American population in their efforts to beat the obesity epidemic and become more healthy. Performing analyses and interviews like these can enable the designers of fitness trackers to create technology that will be more effective and usable, thereby helping users meet their fitness goals and lead healthier lifestyles.

For the Fitbit specifically, the addition of calibration capabilities to increase accuracy over time would greatly improve the long-term use of the device. Giving expert users a way to tune their Fitbit and improve their data quality would encourage continued use of the system as well. Users would also be more understanding of the Fitbit's ability to be "truthful" if they are able to contribute to the effort to make their data more accurate. Finally, an easy addition to improve visibility of system status would be a light for when the device is "tracking" and a different one for "charging needed." These additions to the Fitbit would make it more "honest" about its accuracy and communicative about its current state. Fitbit and other health technology designers should bear in mind that it is the user's intuitive understanding of a product that makes it simple, not only its sleek and minimalist design.

References

Center for Disease Control and Prevention. (2012, January). Adult Obesity Facts. Retrieved from http://www.cdc.gov/obesity/data/adult.html

Fitbit One Wireless Activity and Sleep Tracker Black. (2013). Retrieved from http://www.shopping.com/fitbit-fitbit-one-wireless-activity-and-sleep-tracker-black/info?sb=1

Grice, Paul. (1975). "Logic and conversation". In Cole, P.; Morgan, J. *Syntax and semantics* (pp. 41–58). New York: Academic Press.

Guzman, M. (2013, March 2). Analyze this: Quantified Self is not as geeky as you think. *The Seattle Times.* Retrieved from http://blogs.seattletimes.com/monica-guzman/2013/03/02/analyze-this-quantified-self-is-not-as-geeky-as-you-think/

Guzman, M. (2013, January 28). Using tech to change your habits? Lessons from a behavior change fanatic. The Seattle Times. Retrieved from http://blogs.seattletimes.com/monica-guzman/2013/01/28/using-tech-to-change-your-habits-lessons-from-a-behavior-change-fanatic/

Mackinlay, M. (2013, March 16). Fitbit. Retrieved from http://www.fitbit.com/from/2013/02/17/to/2013/03/16

Neilsen, J. (1995, January 1). 10 Usability Heuristics for User Interface Design. *NN Group*. Retrieved from http://www.nngroup.com/articles/ten-usability-heuristics/

Resnick, M., Maloney, J., Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., & Kafai, Y. (2013). Scratch: Programming for Everyone. *Communications of the ACM*. Retrieved from http://scratch.wiki.hoover.k12.al.us/file/view/scratch-cacm.pdf

Richmond, S. (2013, January 3). Wearable Technology Could Give You a New Year Fitness Boost. *The Telegraph*. http://www.telegraph.co.uk/technology/news/9776667/Wearable-technology-could-give-you-a-New-Year-fitness-boost.html

Schybergson, O. (2013, March 18). Making the Wearable Tech Revolution a Reality. *CNN Money*. Retrieved from http://tech.fortune.cnn.com/2013/01/29/making-wearable-technology-a-reality

Shrager, Jeff. (2013, March 22). Course Glossary for Symsys 145/245: Interaction Analysis. Retrieved from http://jeffshrager.org/symsys/notes.html

Appendix: System Documentation

These are pictures of the web-portal and the iPhone app which the Fitbit connects to. These four pages show data about activities performed, floors climbed, miles traveled, and calories burned. The first picture of the web portal shows a more detailed graph of calories burned, and the second web portal image shows time active versus sedentary. The two iPhone snapshots show activities performed along with the typical activity level view.