When Fitness Trackers Don't 'Fit': End-User Difficulties in the Assessment of Personal Tracking Device Accuracy

Rayoung Yang[1], Eunice Shin[2], Mark W. Newman[1, 3], Mark S. Ackerman[1, 3]

[1]School of Information, [3]Dept. of EECS
University of Michigan, Ann Arbor, MI, USA
{rayang, mwnewman, ackerm}@umich.edu

[2]IBM Interactive Experience
Chicago, IL, USA
eyshin@us.ibm.com

ABSTRACT

Personal tracking technologies allow users to monitor and reflect on their physical activities and fitness. However, users are uncertain about how accurately their devices track their data. In order to better understand this challenge, we analyzed 600 product reviews and conducted 24 interviews with tracking device users. In this paper, we describe what methods users used to assess accuracy of their tracking devices and identify seven problems they encountered. We found that differences in users' expectations, physical characteristics, types of activities and lifestyle led them to have different perceptions of the accuracy of their devices. With the absence of sound mental models and unclear understanding of the concepts of accuracy and experimental controls, users designed faulty tests and came to incorrect conclusions. We propose design recommendations to better support end-users' efforts to assess and improve the accuracy of their tracking devices as required to suit their individual characteristics and purposes.

Author Keywords

Fitness tracker; sensor-based tracking; activity; accuracy.

ACM Classification Keywords

H.5.2 User Interfaces, H.5.m Miscellaneous.

INTRODUCTION

According to a 2013 Pew Research study, 60% of U.S. adults track at least one health indicator such as their weight, diet, or exercise routine [22]. With the availability and popularity of commercial tracking devices such as Fitbit, there has been a rapidly growing interest in sensor-based tracking devices for monitoring personal health and fitness and the number of users is increasing.

Personal tracking devices use various types of sensors to provide data about users' physical activities such as steps taken, distance traveled, and calories burned [23]. Some devices monitor sleep quality and still others add sensors to track biometric data such as heart rate, perspiration, and body temperature (e.g., [24]). With these data, users want to track their activity levels as well as obtain insights into their daily behavior and health status [15].

While the popular reception to these technologies has been primarily positive, a growing chorus of critics caution users to be aware of how accurate or useful tracked data is, and in particular to avoid developing a flawed or incomplete picture about their activity level or health condition (e.g. [25,26]). Usability issues have been cited as a factor that make it difficult for users to test and improve the accuracy of their data [17], however, to address these issues we need a better understanding of how users perceive and assess accuracy. Armed with such an understanding, we can hopefully design more usable and useful devices that support users' needs and practices around assessment.

In this study, we set out to try to understand how important accuracy is for self-trackers and what processes they use to assess the accuracy (i.e., trueness and precision) of data measurements. To do this, we started where many consumers start when deciding whether or not to purchase a particular tracking device—with product reviews. By studying both positive and negative user-contributed reviews of the most popular devices on the market, we were able to understand not only the most common successes and failures, but also the differences in individual motivations and usage patterns that led to both positive and negative outcomes. To help us calibrate our insights from the analysis of reviews, and to help us delve more deeply into the personal reasons behind users' value assessments, we also interviewed 24 people who had used a fitness tracker for at least one month.

The findings of our study offer three contributions to an understanding of how people perceive and assess the accuracy of sensor-based personal tracking devices. First, we found that users' satisfaction with accuracy (trueness and precision) varied because personal tracking devices did not adequately attend to differences between user characteristics (e.g., physical characteristics, bodily

UbiComp '15, September 7–11, 2015, Osaka, Japan.

ACM 978-1-4503-3574-4/15/09...$15.00.
http://dx.doi.org/10.1145/2750858.2804269

movements, activities types, and lifestyle), as well as their purposes and expectations for data accuracy.

Second, while users had expectations of these devices, many did not have a sufficient understanding of how these devices work to detect and measure the phenomena of interest, and thus did not know how to assess data quality or measurement errors. Without a clear understanding of the relevant concepts of accuracy such as trueness, precision, types of errors, or experimental controls, users often conducted faulty tests and came to incorrect conclusions. Third, we identified seven types of problems that users encountered as they sought to assess accuracy of their personal tracking devices.

Based on our findings, we offer design implications for personal tracking devices to better support end-users' efforts to assess accuracy of their data including *supporting testability*, *allowing greater end-user calibration*, and *increasing transparency* to improve the usefulness of sensor-based personal tracking devices.

BACKGROUND AND RELATED WORK

Personal activity trackers [23] are devices worn on the body that help keep track of a various types of activity data about the wearer including steps taken, distance traveled, calories burned, and sleep quality. The most common sensing technologies used in these devices are 3-axis accelerometers, which measure speed and direction, and altimeters to track the elevation of the device. Using these data, tracking devices make inferences to detect steps, running, and sleep quality. Additional sensors used in more advanced devices include skin temperature sensors, optical sensors to infer heart rate by measuring blood flow, and electrodermal response sensors to measure perspiration on the skin surface (e.g., [24]).

User perception and reactions to errors

Several studies have investigated how users perceive accuracy and react to errors that arise when using sensor technologies. In a study of everyday experience with a body scale, Kay et al. [8] examined end users' perceptions of accuracy. They investigated different types of unrealistic expectations people had of their weight data and negative reactions they had to perceived inaccuracies. They found that users' perceptions of accuracy, precision, and resolution were often disconnected from the scales' capabilities. In this paper, we are extending Kay et al. by looking at perceived accuracy of activity-tracking devices. We believe assessing the accuracy of such devices introduces new challenges, as personal-tracking devices produce more data points as they continuously track values such as movement and heart rate.

Consolvo et al. [4] found that there were subtle differences in participants' reactions to different types of errors. When participants gained workout credit for their non-workout activities, they were tolerant of errors. However, they were less tolerant of errors that failed to recognize their workout activities that they believed should have been detected. These errors led some users to favor doing exercises that could be credited, a phenomenon also noted in [7]. These studies indicate that the accuracy of self-tracking devices is an important matter as it affects users' engagement and satisfaction with devices.

While previous work identifies issues regarding users' perceptions and understandings of accuracy of personal sensing devices, only a few studies have examined how users test the accuracy of a device's measurements. For example, Mackinlay [17] described users' difficulties in testing and improving the accuracy of their data, and attributed these difficulties to the limited visibility of the system status, which hinders user efforts to test and calibrate their device. However, we further consider the challenges and problems that users encounter as a result of the methods and processes they use to assess the accuracy of their device and explore how devices can better accommodate users' testing practices. Along similar lines, Choe et al. [3] examined Quantified-Selfers' experiments and noted that their self-experimentation often lacked scientific rigor. However, the purpose of Quantified-Selfer's experiments 'on self' was to test their hypotheses or identify potential correlations in terms of their behavior, and not to assess the accuracy of the device measurement. The latter is the focus of our study.

In summary, we believe designers and developers of personal tracking devices would benefit if there were additional work in understanding how users form their judgments and draw their conclusions on the accuracy of a device's measurements.

METHOD

To better understand methods and processes discretionary users undertake to assess the accuracy of their personal devices over time, we drew upon two sources of data: Amazon product reviews and interviews with users of tracking devices.

Amazon Product Review Data

We began our study by looking at user-generated reviews from a popular ecommerce site—Amazon.com. Given that our interest was in understanding how users make sense of their devices and perceive the accuracy of their tracked data, we reasoned that product reviews based on personal experience would be a good source of data for our study since users post a positive or negative assessment of a device's value.

We chose six devices based on the overlapping but diverse range of features they provide and their popularity in the market: Fitbit One, Fitbit Flex, Jawbone UP24, Basis B1, Withings Pulse, and the Polar H7 heart rate sensor. The devices varied in form factor: three devices (Flex, UP24, and B1) were wristbands, Fitbit One and Withings Pulse were clip-ons with a sleep wristband, and Polar H7 was a chest strap.

Given the large number of reviews for each device (18,772 across all six devices), we chose to analyze the top 100 "most helpful" reviews for each device (using Amazon's sort order), resulting in a data set containing a total of 600 reviews. Based on an initial read of the product reviews, we determined that the "most helpful" reviews were generally lengthy, well-written, and rich in personal detail, thus providing a valuable data set for analysis. Moreover, the "most helpful" reviews contained a balance of both positive and negative reviews.

Interview Data

While product reviews offer a good overview of the range of problems and benefits offered by the devices, Amazon reviewers may not represent typical users, and their stories may not represent common use cases. To help us calibrate our insights from the analysis of reviews, and also to flesh out a more nuanced understanding of the experience of using tracking devices, we interviewed 24 participants who had used a fitness tracker for at least one month. Each interview lasted about 45 minutes and each participant was compensated $20. The interviews were audio-recorded and transcribed.

We recruited participants via various methods including recruiting emails, social networking sites, and contacting individuals who publicly posted about their experiences with fitness tracking devices. We selected the participants based on the type of device they used and the duration of their usage. We sought both current users and individuals who had stopped using their devices. We interviewed 13 females and 11 males. Nine participants had stopped using their trackers for various reasons including motivation loss, reduced curiosity/novelty, inaccurate data, equipment failure, and life or health changes such as pregnancy. They came from a variety of professions such as designer, librarian, financial analyst, student, and fundraiser. Table 1 shows demographic, device ownership, and duration of usage.

Our interviews were semi-structured, informed by the initial findings from the Amazon product review data. In each interview, we first asked each participant about motivations for acquiring the device, followed by detailed questions around device usage, mental models about how the device worked, and experiences with assessing the accuracy of the device.

Data Analysis

First, we analyzed the 600 product reviews to understand why users came to a positive or negative assessment of a device's value, as well as how they made sense of their tracking device and their data. We used open coding to identify recurrent concepts and themes, generating a first-round coding scheme. Second, in coding our interview data, we both used the tentative coding scheme from our product review analysis and actively looked for additional themes that emerged from the interview data. We generated a second-round coding scheme after analysis of the interview transcripts. Both the first-round and the second-round codes were discussed among the research team, and further codes were determined to refine our coding scheme. The process was iterative, as we generated, refined, and probed the themes that emerged from each dataset. Our final coding scheme, then, combined the themes that emerged from both the product reviews and interviews. As a third step, we re-analyzed all product reviews and interview data using the final coding scheme.

ID	Gender	Age	Device	Months of use
1	Female	30	LifeTrak	1
2	Female	28	Nike+FuelBand	12
3	Female	30	Fitbit Flex	9
4	Male	N/A	Basis B1, Garmin FR70	7
5	Male	24	Jawbone UP24	14
6	Female	23	Jawbone UP24	6
7	Male	31	Nike+FuelBand	18
8	Male	23	Basis B1	11
9	Female	21	Jawbone UP24	7
10	Male	40	Fitbit Flex	4
11	Female	37	Fitbit One	1
12	Male	27	Fitbit One	13
13	Female	50	Polar Loop	4
14	Female	27	Fitbit Flex	10
15	Female	25	Fitbit Flex	2.5
16	Female	23	Fitbit Force	3
17	Female	N/A	Jawbone UP24	5
18	Female	29	Fitbit Flex	12
19	Male	27	Basis B1, Fitbit One	18
20	Male	N/A	Basis B1	2
21	Male	72	Fitbit One	12
22	Female	27	Fitbit Flex, Garmin FR	12
23	Male	26	Fitbit Flex, Basis B1	2
24	Male	37	Fitbit Flex	1

Table 1. Summary of interview participants' information

Amazon product reviews were useful as they provided a large amount of information about how individuals thought about and discussed their devices as well as a more diverse range of device evaluation. With our interviews, we aimed to verify what we observed in the reviews and obtain more targeted and nuanced data that complemented, reinforced, and challenged many of the issues we had identified. Therefore, in this paper we present our findings from both datasets together.

In what follows, the term 'users' refers both to Amazon reviewers and interview participants. When presenting quotes and stories, we use "R" followed by the device name and a review number to indicate Amazon reviewers (e.g. R-DeviceName-1) and "P" followed by a participant ID for interview participants (e.g. P2).

FINDINGS

Concerns with 'accuracy' were prevalent in the Amazon review data, appearing in over a third of the analyzed reviews (222 out of 600), according to our coding scheme. Echoing what was reported in [8], people showed a misunderstanding of statistical terms involved in measurement/sensing error. Both reviews and interviewees in our study often used 'accuracy' as an umbrella term to refer interchangeably to both 'trueness' and 'precision'. Before we describe our findings, we clarify our use of key terms using definitions from [18,27].

Accuracy refers to "the closeness of agreement between a test result and the accepted reference value [27]." Two terms, trueness and precision, describe aspects of the accuracy of a measurement method. "***Trueness*** refers to the closeness of agreement between the average value of a large number of test results and the true or accepted reference value [27]". ***Systematic error*** affects trueness, as it tends to shift all measurements in a systematic way [18]. For example, systematic error occurs when an imperfectly calibrated measurement instrument is used, such as with distance measurements by an inaccurately marked meter stick or time measurements by a clock that runs too fast or slow. ***Precision*** refers to the closeness of agreement between test results and "does not relate to the true value or the specified value [27]." ***Random error*** principally affects precision and varies in an unpredictable way. It can occur for a variety of reasons such as noise in the measurement or small sample size.

The rest of findings are organized as follows: First, we describe how users in our study had varying expectations regarding the trueness and precision of data. Second, we describe different methods and processes users used to assess the accuracy of the data through their daily use. Finally, we explain what problems existed in the ways that users conducted their testing processes.

VARYING EXPECTATIONS FOR ACCURACY

In our study, users had varying expectations regarding the accuracy of devices' measurements, echoing previous studies of users of body scales [8] and fitness trackers [7,19]. Building on these findings, we argue that better understanding specific purposes and needs that users have regarding different aspects of accuracy, precision and trueness, provides a rich picture of how users actually perceive the accuracy of the data.

Precision is more important than trueness for trends.

Most users who valued the trends or pattern of their data over time put more importance on the precision. As long as a device was consistent in measuring the phenomena of interest, it was sufficient enough for them to see if they were making progress towards their goals.

From what I can tell, it is fairly accurate. I'm sure it doesn't capture my movement or sleep completely accurately, but it is consistent. In other words, even if it isn't 100% accurate, the trending is directionally correct (for the most part ...). (R-UP-32)

*But, in some cases, accuracy **does** matter.*

For users who wanted to optimize their exercise or track health conditions, it was not acceptable for their devices to make a rough approximation of the correct measurement. Some individuals wanted to use heart rate information to optimize their exercise, such as maintaining their heart rate at certain levels (high and low) during interval training. Other individuals had health issues, such as P1, who had had heart surgery a few years prior. She got her fitness tracker to monitor her heart rate. If P1 felt a 'weird rhythm' with her heart or was stressed, she would check her heart rate on her device to keep watch on her condition. Because P1 used the heart rate measurement for making decisions on whether to seek treatment, the accuracy of the heart measurement was important. She assessed its accuracy by comparing her device's measurements against the measurements taken at her doctor's office.

Lifestyle differences affect accuracy.

Depending on their lifestyles and activities that they did throughout the day, users can come to different conclusions about the accuracy of their device. In our study, some users found their device to be accurate enough for their purposes while others found their device useless. For example, one reviewer found it problematic when her/his device was counting steps when s/he was rocking one's baby, but not when pushing a stroller.

The Flex does not count steps if you are pushing a stroller or cart. This may not be an issue to some, but it is to me b/c I have stroller age children. Then I noticed it was logging hundreds of steps while I was rocking my baby. [...] Maybe something else out there will be more accurate for my lifestyle. (R-Flex-31)

Since there were variances in individuals' characteristics, one device that worked fine for one user might work poorly for another user depending on the individual's purpose or value for the data, or the types of activities they did throughout the day. Different users' expectations or needs with respect to the accuracy varied and led them to conclude that the same result was or was not good enough for them.

HOW PEOPLE ASSESS ACCURACY: METHODS AND PROCESSES

Based on our analysis, many users were curious about the accuracy and reliability of the data and tried to find ways to assess it. Commonly, users tested their device to confirm or disconfirm expectations or hypotheses about the inner workings of the device. Although most users in our study might not have actually known how to systematically test the accuracy of their device measurements, many went through some kind of process, whether consciously or not, to assess the accuracy and reliability of the data.

In this section, we describe two types of process that we observed from our study: 1) ad-hoc assessment and 2) 'folk-testing'. *Ad-hoc assessment* describes situations that involve few instances of measurement results, and rely on one's intuition than other more objective means. With '*folk-testing*', we describe more systematic testing that involves comparing measurement data with a more reliable source or crosschecking with other devices.

Ad-hoc Assessment

Most commonly, users in our study assessed the accuracy of their device in a natural, intuitive manner. They found quick and convenient ways to verify their expectations.

I do believe that it works fairly accurately as a pedometer. Though I can't say I actually counted my steps in order to test it, I did note that on days in which I walked a lot, it seemed to register a lot more steps than on days in which I did not. (R-Basis B1-83)

Over time users did a variety of activities and encountered or noticed on certain occasions where they found that the data was not as accurate as they believed. Then, they began to suspect that the device was not accurate when presented with data that violated their expectations.

I did figure out irratic [sic] arm movement gets credited as steps. I washed 11 puppies this past weekend and it looked like I walked 12,000 steps that day when really it was the back and forth arm motion of washing the dogs. (R-Flex-42)

Another common type of ad-hoc assessment appeared when users tried to confirm or disconfirm their expectations [10]. For example, they looked for cases when the device correctly acknowledged a movement that was supposed to be detected. Or, they intentionally made movements that should not be detected to check whether the device got confused and detected them incorrectly.

I tried an experiment. I walked briskly for 30 minutes without moving my arms. Then I did arm exercises sitting on the couch for 30 minutes. The band recorded the arm movements as exercise and did not record the walking as activity. That is a MAJOR FLAW in the band! (R-UP-60)

Folk-testing

While many users used ad-hoc assessment to form an impression of their device accuracy through daily use, a subset of users engaged in more structured testing to try to characterize the device's data accuracy. However, most users did not really know how to validate the accuracy since they had limited understanding of various sensors and algorithms used in their fitness trackers.

In the absence of formal training or documentation on how a device works, people often developed "folk theories" of devices to make sense of how the device worked and thus operate the device according to their needs [9]. We observed that users in our study developed folk theories about the inner workings of the devices such as sensors and algorithms. They came up with various ways to verify how the device worked as well as to assess the accuracy of the data that their device provides. We call this assessment process 'folk-testing' to draw an analogy to 'folk theory' as it shares characteristics. Folk-testing is 1) inconsistent when compared to an expert's model/practice, 2) "acquired from everyday experience," 3) sufficiently functional for everyday use, and 4) "varies among individuals, although important elements are shared" [9].

In this section, we describe what methods and processes users used to conduct testing in a more deliberate and systematic manner. However, the extent of the systematicity was varied.

Comparison with "ground truth"

In several cases, testing was carried out in a systemic way by comparing the device output with a measurement made by the user using a method perceived as accurate and reliable. In fact, approximately one-fifth of the reviews that mentioned accuracy (46 reviews, according to our coding scheme) described the use of testing methods to compare with other commercial devices. For example, the majority of users manually counted their steps, then compared their tally to the number counted by the device:

I've done a couple quick tests, where I've walked around the building with the app up and I watched the step count increase as I'm actually taking steps. It's pretty close. (P10)

Another user used a similar method, but found that the device was repeatedly counting only half of the steps.

I walked from my office at work all the way to the other end of the building and back again, counting my steps. This device only recorded about half of my steps. I did this several times a day throughout the day, with the same results. (R-Pulse-65)

P15 concluded that his device was accurate because the distance measurement was close to the distance measured using distance markers posted along the path he ran.

The path I ran […] had stakes permanently planted to show you the distance you had covered. Of course anything could be slightly off in that way, but it was close to what had estimated you had run at least, around this path. Which I think is a fairly good measure. (P15)

In another case, one reviewer used a similar method and found the device measurement was not accurate, when compared to a reliable measurement:

I run on a track at the local college. When I run two miles, Pulse says I ran three. When I run 4 miles, Pulse says I ran seven. When I run 6 miles, Pulse says I ran 11. This amount of error makes the device useless for any kind of fitness tracking (R-Pulse-45).

The cases above demonstrate that even though users used a similar method, they came to different conclusions about the accuracy of the data.

Comparison with other commercial devices

Many users crosschecked with other measurements to assess the accuracy of their device. Approximately one-fourth of the reviews that mentioned accuracy (58 reviews, according to our coding scheme) described their testing methods to compare with other commercial devices. One reviewer conducted an exceptionally thorough testing:

> **TEST #1: Normal stride with my FWA [Fitbit Wearing Arm] as relaxed as possible (i.e., just swinging at my side): Distance traveled according to Endomondo [a mobile app]: 1.78 miles Steps taken according to Fitbit: 3,564 Average steps per mile: 1.898.87 *TEST #2: Normal stride, while holding and regularly sipping a coffee in my FWA: Distance […]: 1.80 miles. […] Fitbit: 3,393 Average steps per mile: 1,885.00 Wow! I'm pretty freaking impressed. The difference between my normal stride and my coffee holding/sipping stride was only 4 steps per mile! [...] *TEST #3: I have a Stamina In-Motion Elliptical Trainer that counts your "steps" as you use it. The elliptical counted 4,000 steps and the Flex counted 3,975. I think a difference of only 25 steps is more than great. (R-Flex-10)*

Another reviewer had also tested three different fitness devices to decide which was worth the trouble to wear. Because of the significant discrepancy among the devices, s/he concluded that none of them was trustworthy:

> *I wouldn't trust any of them to measure […] correctly. Take today - it's 8 p.m. on a typical day. I swam 1.25 miles this morning, then came to the office and mostly sat at my desk. Here are my steps: Polar Loop: 12,819, no miles measured… Fitbit One: 5,203 steps; 2.56 miles […] Withings Pulse: 4,336 steps, 2.09 miles […] So, which one would you trust? Who the heck knows? (R-Pulse- 96)*

As noted earlier, similar testing methods resulted in opposing test outcomes. A user could find their device to be very accurate, completely random, or way off. This discrepancy might be due to 1) a device that was defective, 2) the way users designed or executed the testing, or 3) the underlying models used by the devices fitting different individuals differently.

PROBLEMS IN AD-HOC ASSESSMENT AND FOLK-TESTING

In the previous section, we described approaches that users in our study used to assess accuracy of their device. While the majority of users were comfortable with assessing accuracy in an ad-hoc way, some users made more effort to plan more systematic tests.

However, in both we found that it was often difficult for users to assess whether the data was accurate, or if not, why the device gave the inaccurate measurements, or what types of error were affecting the accuracy. In this section, we describe problems with ad-hoc assessments and folk-testing.

Problem 1. Users conducted testing in uncontrolled or incomparable conditions.

We observed that several users in our study made the wrong assessment about accuracy by comparing measurement counts in different physical conditions. For example, P6 went hiking with her boyfriend, who had the same Jawbone UP, and compared her step count to his. As P6 found it was significantly off, "*at least 3,000 steps,"* she concluded that her device was not reliable, and tried to come up with reasons to explain the difference:

> *He was carrying a backpack around him and he kept adjusting his backpack. So, it was maybe that added to the steps. Then, he had his camera in his hand. He kept taking pictures, and maybe it added that too as a step. (P6)*

While P6 could rationalize the differences, it is important to understand that her rationale was disconnected from the statistical concepts. When P6 encountered discrepancy between her and her boyfriend's step counts, they were trying to use inter-device reliability of two devices with unknown accuracy as a proxy for accuracy. Moreover, they were judging two different devices based on their estimations of a measure specific to its wearer – steps. This is a case where differences in step count could be attributed to differences between properties of the individuals, such as stride length, rather than the accuracy of either device. Alternatively, the problem could have been fit between the individual and the device's model of that individual.

When testing accuracy of data measurements, it is problematic to compare different measurement data in uncontrolled and incomparable conditions. However, users often conducted their testing in such conditions. With absence of a clear understanding of the relevant concepts of accuracy, trueness, precision and experimental controls, users might design faulty tests and/or come to incorrect conclusions.

Problem 2. Different devices have different models, but there is not a standard model to compare.

If users had more than one device, they were likely to assess the accuracy, by comparing between or amongst multiple devices. However, there are limitations in this method in that it lacks an absolute standard, making it hard to resolve discrepancies. Different devices employ different definitions of the phenomena they are measuring and have different models and algorithms to measure or detect the phenomena.

When people design a set of tests to compare multiple readings of devices in different conditions, it is difficult to know which device is accurate and precise among others without having a reliable truth measurement. For example, as previously mentioned, R-Pulse-96 wore three different fitness devices, Polar Loop, Fitbit One, and Withings Pulse, to compare measurements against each other. However, s/he was not able to tell which of those three devices was accurate. It could be argued that users needed a more reliable device or measurement that had been proven to be accurate in order to make a judgment regarding the accuracy of their device measurement. However, as seen in our study, users oftentimes did not have an access to such a device. Thus, users tended to rely on their perception.

Problem 3. Users are not aware of systematic error or how to address it.

Often users did not understand that the algorithm models that different devices used for calculating the steps might be the cause of the discrepancy between measurements. For example, one reviewer compared measurements between Fitbit One (worn on the hip) and Fitbit Flex (wore on the wrist), and found Fitbit Flex's measurement was higher than Fitbit One. This user felt one's Fitbit Flex was not accurate for her/his purposes.

I compared the Flex's count with the One's (which I've found very accurate)… Flex dramatically over-estimated the step count, by about 20%. That's not helpful when you're trying to get fit. (R-Flex-37)

Different people may fit the underlying models used by the devices better, resulting in systematic biases. This could be an explanation for the issue described above by R-Flex-37. In case of the device being consistent but not accurate, the measurement error was likely due to systematic error. Although calibration (e.g., entering one's stride length information to the device) could diminish systematic errors in measurements, oftentimes users did not know the benefits of such calibration and might abandon the device rather than try to make it work for them. Another user (R-Flex-27) had a similar problem in that the device was consistently over-rating her/his activity. This user did not know that calibration could solve the issue, and ended up returning the device.

Problem 4. Test cases do not replicate a realistic scenario.

Many users suspected that their device was tracking many other types of movement besides "steps." Some users tested with different movements to see which of their movements would get counted and which would cause erroneous tracking:

I tried jumping. I tried punching. I tried just swinging around. I tried tapping it on things. I tried a whole bunch of stuff just to see if I could get false steps and it was pretty difficult. There's a very specific punch, that you would really have to go out of your way to do, and it would occasionally count a steps for it. Other than that, it really didn't count anything that wasn't activity. (P8)

I actually tried to see if I could manipulate the movement just to see whether it would track my hand waving up and down as a step and just to see how sensitive it is to different types of things. […] I start really small and just get larger and larger just to walk through a set of patterns […] which I would really can't say […] would be my [normal] types of movements. (P19)

These test cases that users in our study came up with were often aimed at testing the sensitivity of the device—i.e., whether the device could accurately pick up on certain movements. However, the sequence of the movements that users tested were not likely to be the ordinary movements they would make on daily basis.

Problem 5: The black-box nature of device algorithms inhibits users from understanding how it works.

There were many problems related to end-users' not understanding how the device worked. In the absence of a coherent mental model of how their device senses and processes sensor data, it was difficult for users to improve the accuracy or the usefulness of their device. Several users explicitly asked for more clear and detailed explanations about how their device worked to measure data or what the data actually meant, and found the lack of such explanations condescending:

I find Jawbone's explanation that advanced algorithms are used to be condescending. Like telling a child that something is too complicated for them to understand. […] I do wish that Jawbone would elaborate about how sleep and activities are tracked so that I can better understand what it is tracking. (R-Jawbone-53)

One factor that contributed to the difficulty was lack of understanding of how the devices calculated the data. For example, a number of users did not understand how "calories burned" data was calculated. Several were confused and expressed frustration: *"What am I doing wrong? It counted my steps fine, but it had me burning tons of calories and I've barely been out of bed. Not so sure about this device after all. (R-Basis B1-70)"*

Most users found it even more difficult to assess these higher-level inferred values (e.g., calories) than they did the accuracy of measurement. One reviewer found that when s/he was playing a video game, her/his device detected the activity as sleeping. Based on this incident, this user concluded that the device used the time of day for sleep tracking. While it might have been the lack of movement that caused the inaccurate inference, this user considered it to be a shortcoming of the device.

I cannot vouch for accuracy [of sleep data …] since I am unconscious. Once again, there is some wizardry going on to determine the levels of sleep since it isn't connected to your head, eyes, or even your lungs. I know some of it is time-based, which brings me to one of the shortcomings – I was playing a video game for 4 hours and it thought I was sleeping most of the time. (R-Basis B1-78)

When users noticed errors in their data measurements, most of them sought to understand the circumstances that led to the inaccuracy. However, it was difficult for them to understand why problems happened or to improve accuracy.

I went to the gym and did over 30 minutes on an exercise bike and another 45 minutes lifting weights. The tracker only showed 2 minutes of very active minutes. I didn't go to the gym or do anything physical the day before and the day after yet somehow ended up with more very active minutes. How?!?! (R-Flex-59)

Problem 6: The underlying phenomena being measured are often imprecisely defined.

Many interview participants commented that there was not a clear definition of measurements such as 'a step', 'sleep', and 'calories burned.' They noted that individuals might have different interpretations. They were uncertain about how their device actually defined the events that it measured.

[Basis-B1] knows when I've kind of tilted my body enough to do a step or two. But, if I literally just stand up, [do] I instantly get the half one step or partial step? I'm not sure how it actually determines the full step. (P19)

I do find the "times awaken" deeply flawed. My unit tells me I was awakened 14x, 9x, etc. I find that excessive and inaccurate. But then again, how does the Fitbit calculate "awaken" activity? I don't know." (R-Jawbone-8)

While all users did not have a clear understanding of the logic behind how their device measured their activities, some users came up with their own justifications:

It doesn't count your steps until you have taken several steps (I think around 15) before it would start counting. At first I found it unfair but then I realized this is for my health and the steps that I want it to really count are the ones where I really am being active and not just walking to my kitchen. (R-Basis B1-7)

All users who had some understanding of how their devices worked or what factors were used to calculate the data, they tended to be more tolerant of the inaccuracy, acknowledging the limitations:

The key really is to have the appropriate expectations of what a device like this is going to do for you. It's a wrist-based pedometer, and that means that if you're doing any activity that over or under involves your hands/arms, it's going to throw off the accuracy of the step counting, regardless of how good Fitbit's accelerometer algorithms are. (R-Flex-13)

Problem 7: Folk-testing misguides users' efforts to improve accuracy.

P2 tested her device while making different hand movements and concluded that her FuelBand detected the speed and magnitude of movement. However, she later found that her understanding was not entirely correct.

As soon as I bought it, I walked with my arms steady vs. swinging it widely. I found that the number would go up faster as I walked with my arms swinging more rapidly and widely. I've tested this numerous times. [...] That's why I figured to wear it in my ankle when I was biking. [...] I was disappointed that wearing it on my ankle did not detect a bike ride. (P2)

She wore her FuelBand on her ankle during bike rides instead of her hand because her ankle was moving while her hands stayed still holding on to the handles. That was a logical workaround based on her understanding; however, she found that the device did not track the motion associated with bike riding. In the end she was left with no insight into how to make the device work for her.

A relatively small but interesting group of users attempted to avoid 'messing up' the readings and tried to make their movements easier for the device to detect. For example, users mentioned a conscious effort to improve accuracy, such as swinging their arms while working or avoiding holding a leash or a bag with the hand wearing the device. However, users in our study preferred their device to detect their natural movements. P2 mentioned that she did not like that she had to make a conscious effort, as it felt 'unnatural' for her to keep paying attention to her body movements for accurate measurement.

DISCUSSION

In this section, we first discuss why we need to understand and support people's methods for assessing the accuracy of personal tracking devices. Then we propose recommendations to better support users in their process to assess quality of measurement data in their device.

One seemingly obvious solution is to build personal tracking devices that provide higher reliability. However, this solution is not straightforward. There are already medical-grade tracking devices that provide much higher reliability than widely available devices. They are not made by consumer electronics companies, not fashionable, not easy to use or comfortable to wear, and much more expensive. The very factors that have made personal tracking devices popular among mass-market consumers (low-cost, comfort, convenience, and design that prioritizes fashion) are the ones that weigh against optimizing for reliability. The reliability and quality of data that personal tracking devices provide may increase over time, but limitations of technical capabilities and users' lack of ability to address those issues are going to be a part of the user experience for the foreseeable future.

We propose design implications that enable users to take better advantage of personal tracking devices. We based our implications on key findings in our paper:

These devices will work differently for different people. This could be because people walk differently, wear devices differently, and have different characteristics (e.g., physical dimensions, bodily movements, activities types, and lifestyle). Moreover, people's expectations or needs with respect to trueness versus precision may lead them to decide that the same result is or is not good enough for them. Thus, *everyone* will need to be able to understand and assess the accuracy and error characteristics of the device well-enough *for their particular characteristics and purposes*, unless the device happens to work well enough for their purposes out of the box.

Conventional approaches to the recognition of human activities using sensor data through classification models struggle to cope with the population diversity problem [13]. Many devices use generalized models and algorithms that are designed to work well for the population norm. While the majority of users might not care about the reliability of the device, in our study, we observed that users were confused and frustrated when they were uncertain about how their device was tracking their data, and many attempted to systematically evaluate through (folk-testing) the accuracy of a device.

We observed cases where the way users designed or executed the testing was problematic and conclusions were made based on poorly designed tests. Users' understanding of the concepts of accuracy, trueness, precision, and types of errors were also disconnected from the underlying statistical concepts. Based on the problems and challenges that we observed in the way users struggled to assess the accuracy of their devices as well as to understand how the devices work, we conclude that it is critical to be attentive of how users currently perceive and assess the accuracy of their data and determine the reliability of their devices. Accordingly, we provide three design implications: *support testability*, *allow greater end-user calibration*, and *increase transparency*.

Support testability

Each user employs their devices under a different set of conditions and therefore presents different considerations with respect to precision, i.e., closeness of agreement between independent measurements of a quantity under the *same conditions* [18]. However, every user has a *different set of conditions*, and thus they have a different measure of precision. This means that end-users have to understand their particular *conditions* as they assess precision. However, due to the limited knowledge acquired from using the device (or a different but similar device), we observed that users sometimes adopted wrong assumptions to guide their attempts to test the precision of the data. This is likely why reviews contradict each other even when they were using a similar method, and why one person's testing procedure might not work for someone else.

For most of the users in our study, we found ad-hoc assessment or folk-testing confirmed or disconfirmed users' mental models, hypotheses, or assumptions in part, but did not test them thoroughly. Users' testing methods helped users to reinforce or disprove their hypotheses regarding their expectations on what could be tracked. However, such confirmation or rejection was, of necessity, based on a limited set of tests. It did not necessarily help users to understand what to predict for unknown or untested scenarios.

Given the limited set of test cases attempted by most users, users need support when figuring out how to select the situations that are most important to them and assess the device functionality in those cases. Therefore, it will be significantly important to provide users with guidelines on how to test their devices for accuracy. This will help to increase the breadth and depth of testing cases, which will lead to a better understanding of the true limitations of the device. Guidance on testing can help alleviate the problems of folk-testing by providing more complete and appropriate set of testing cases to help users to have more realistic and accurate expectations.

Allow Greater End-user Calibration of Models

Earlier we showed several examples where users did not take advantage of calibration features. Individuals have different patterns of movements, and they engage in different types of activities with varying levels of frequency. Lack of capability to manipulate the sensing sensitivity led users to develop workarounds for various and changing daily routines and activities. Some adaptations were effective at delivering value to the user, while others were not.

If a device is designed to track long-term data about oneself, and is meant to be tracking such data continuously over a long period of time (even for a life time), users are likely to expect to be able to customize the device to better fit their individual needs. One could argue that allowing users to do the calibration is the best approach: e.g., devices could "learn" based on humans providing them with information, such as how to accurately detect what needs to be detected (e.g., "steps").

In addition to making it easier to find, enter, and revise calibration settings such as "stride length," we suggest allowing for ways to calibrate the device to personal movement patterns and purposes. Users should be able to record unique movements into a device to let the device know which movements to record and which to ignore. For example, one user might want to count walking around the house while another might not. One might want to indicate not to count when s/he pats a baby. Such examples could provide positive and negative examples to allow a device to learn personalized models of movement, sleep, or other activities, though research will be required to learn the limits of such personalization.

The process for lay-users to train and debug models/classifiers can be better supported by an interface that allows users to provide feedback to a machine learning system for debugging that system. For example, Kuleza et al. [11] proposed an approach to help end users effectively and efficiently correct shortcomings and personalize machine learning systems.

Designers of personal tracking devices can take advantage of greater end-users' calibration of models, not only to increase accuracy for individuals, but also to better address diversity in user populations [13]. For example, Lane et al. [13] proposed Community Similarity Networks, which exploits crowd-sourced sensor-data to personalize classifiers with data contributed from other similar users.

Increase transparency

From this study, we learned that very few users of personal tracking devices had a solid understanding of how sensors work to track their movement or heart rate. For example, users wondered, "What is considered a step for the device and how does it determine it?" Users figured out that step counting had to do with motion sensing, but not to the extent that they understood what types of motion the device could measure. Another problem was that the limitation of testing methods and processes that users used and the results of such testing might suggest misguided workarounds to users.

Often, users expressed frustration when the device was misdiagnosing their movements. This frustration often led to dissatisfaction and further disengagement (as also seen in [5,17,20]). In order to reduce users' frustration, designers should avoid misleading users into developing wrong or insufficient mental models, but provide a clear view of what the device can deliver and how. As an example of providing such transparency, a device could show visualizations of movement patterns of its user for a period of time with explanation of how the device views those movement patterns (steps, walking up the stairs, running, etc.). Since a person knows what movement they made for that time period, this can give a good depiction of how a device's accelerometers sense that person's movements.

When presenting higher-level inferences with conceptual definitions, such as calories burned or sleep quality, personal tracking devices should provide more support for users to better understand the definition of the phenomenon being measured and how the device measures them. It was perhaps more frustrating when users found higher-level inferences were not accurate since it is more difficult or even impossible for users to assess the accuracy of calories or sleep quality than steps counts and distance measurement.

When a system is working properly, as expected by its users, knowing 'how-to-use-it' may suffice [6]. However, when a system is behaving in an unexpected and erroneous way, a user's understanding of 'how-it-works' becomes more crucial for the user to be able to identify errors and fix the problems [6]. As described earlier, users were confused and frustrated when they found their sleep data was not accurate but were not able to address the issue.

This highlights a specific challenge for users' mental models of the systems that employ sensing and algorithms. While a user needs a more detailed and deeper understanding of the system's logic or reasoning, its components of processes, such as sensing, inference and machine learning techniques are concealed and difficult to understand for a novice user [6,14,21]. Despite the importance of mental models, which have been studied in depth, there is a lack of research investigating the effect and development of a user's mental model of sensing systems. Such research [1,2,12,16] will be important for helping people attain the desired value from personal tracking technologies.

CONCLUSION

In this study, we analyzed 600 Amazon reviews and 24 interviews to better understand how users engaged with and adapted to self-tracking technologies. From the primary findings of our study, we conclude that it is critical to be attentive to how users currently perceive and assess the accuracy of their data to determine the reliability of the devices. By supporting more robust testability, increased transparency of what devices can and cannot detect, and allowing for ways to personalize models of movement, sleep, or other activities, personal trackers will be able to enhance and furthermore transform the lives of many people.

ACKNOWLEDGMENTS

This work was funded in part by NSF grant IIS-1149601. We thank our participants and the anonymous reviewers for their insightful feedback that improved the paper. We would also like to thank Jodi Forlizzi, Tawanna Dillahunt, Emily Provost, and the Interaction Ecologies group for providing valuable feedback.

REFERENCES

1. Victoria Bellotti, Maribeth Back, W. Keith Edwards, Rebecca E. Grinter, Austin Henderson, and Cristina Lopes. 2002. Making sense of sensing systems: five questions for designers and researchers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02), 415–422. http://dl.acm.org/citation.cfm?id=503450
2. Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4, 193–212.
3. Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 1143–1152. http://dl.acm.org/citation.cfm?id=2557372

4. Sunny Consolvo, David W. McDonald, Tammy Toscos, et al. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08) 1797–1806. http://dl.acm.org/citation.cfm?id=1357335

5. Cara Bailey Fausset, Tracy L. Mitzner, Chandler E. Price, Brian D. Jones, Brad W. Fain, and Wendy A. Rogers. 2013. Older Adults' Use of and Attitudes toward Activity Monitoring Technologies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications, 1683–1687. http://pro.sagepub.com/content/57/1/1683.short

6. Robert M. Fein, Gary M. Olson, and Judith S. Olson. 1993. A mental model can help with learning to operate a complex device. *INTERACT'93 and CHI'93 conference companion on Human factors in computing systems*, 157–158. http://dl.acm.org/citation.cfm?id=260170

7. Thomas Fritz, Elaine M. Huang, Gail C. Murphy, and Thomas Zimmermann. 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* (CHI '14), 487–496. http://dl.acm.org/citation.cfm?id=2557383

8. Matthew Kay, Dan Morris, Julie A. Kientz, and others. 2013. There's no such thing as gaining a pound: reconsidering the bathroom scale user interface. In *Proceedings of the ACM international joint conference on Pervasive and ubiquitous computing* (UbiComp '13), 401–410. http://dl.acm.org/citation.cfm?id=2493456

9. Willett Kempton. 1986. Two Theories of Home Heat Control*. *Cognitive Science* 10, 1, 75–90.

10. Joshua Klayman and Young-Won Ha. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review* 94, 2, 211.

11. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 126-137. http://doi.acm.org/10.1145/2678025.2701399

12. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1–10. http://dl.acm.org/citation.cfm?id=2207678

13. Nicholas D. Lane, Ye Xu, Hong Lu, et al. 2014. Community Similarity Networks. *Personal Ubiquitous Comput.* 18, 2 (February 2014), 355-368. http://dx.doi.org/10.1007/s00779-013-0655-1

14. Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '15), 1603–1612. http://doi.acm.org/10.1145/2702123.2702548

15. Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the international conference on Ubiquitous computing (UbiComp '11)*, 405–414. http://dl.acm.org/citation.cfm?id=2030166

16. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 2119–2128. http://dl.acm.org/citation.cfm?id=1519023

17. Molly Mackinlay. 2013. Phases of Accuracy Diagnosis:(In) visibility of System Status in the Fitbit. *Intersect: The Stanford Journal of Science, Technology and Society* 6, 2. http://ojs.stanford.edu/ojs/index.php/intersect/article/view/555

18. Antonio Menditto, Marina Patriarca, and Bertil Magnusson. 2007. Understanding the meaning of accuracy, trueness and precision. *Accreditation and quality assurance* 12, 1, 45–47.

19. John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers Chalmers. 2014. Personal tracking as lived informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 1163–1172. http://dl.acm.org/citation.cfm?id=2557039

20. Patrick C. Shih, Kyungsik Han, Erika Shehan Poole, Mary Beth Rosson, and John M. Carroll. 2015. Use and Adoption Challenges of Wearable Activity Trackers. *Proc. iConference*. http://hdl.handle.net/2142/73649

21. Rayoung Yang and Mark W. Newman. 2013. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the international conference on Ubiquitous computing (UbiComp '13)*, 93–102. http://dl.acm.org/citation.cfm?id=2493489

22. Main Report | Pew Research Center's Internet & American Life Project. Retrieved February 21, 2015 from http://www.pewinternet.org/2013/01/28/main-report-8/#fn-87-3

23. Fitness Tracker Comparison Chart - Best Fitness Tracker Reviews. Retrieved February 21, 2015 from http://www.bestfitnesstrackerreviews.com/comparison-chart.html

24. Basis | The Ultimate Fitness and Sleep Tracker. Retrieved February 21, 2015 from https://www.mybasis.com/

25. For Wearables, Accurate Sensing Is Tricky | MIT Technology Review. Retrieved June 23, 2015 from http://www.technologyreview.com/review/538416/the-struggle-for-accurate-measurements-on-your-wrist/

26. Sleep-tracking gadgets raise awareness - and skepticism. Retrieved February 21, 2015 http://www.usatoday.com/story/news/nation/2013/03/24/sleep-tracking-devices/2007085/

27. ISO 5725-1:1994(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions. Retrieved July 1, 2015 from https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en