

There's No Such Thing as Gaining a Pound: Reconsidering the Bathroom Scale User Interface

Matthew Kay,^{*†} Dan Morris,^{*} mc schraefel,^{*‡} Julie A. Kientz[†]

^{*}Microsoft Research
Redmond, WA, USA

[†]University of Washington
Seattle, WA, USA

[‡]University of Southampton
Southampton, UK

mjskay@uw.edu, dan@microsoft.com, mc@ecs.soton.ac.uk, jkientz@uw.edu

ABSTRACT

The weight scale is perhaps the most ubiquitous health sensor of all and is important to many health and lifestyle decisions, but its fundamental interface—a single numerical estimate of a person's current weight—has remained largely unchanged for 100 years. An opportunity exists to impact public health by re-considering this pervasive interface. Toward that end, we investigated the correspondence between consumers' perceptions of weight data and the realities of weight fluctuation. Through an analysis of online product reviews, a journaling study on weight fluctuations, expert interviews, and a large-scale survey of scale users, we found that consumers' perception of weight scale behavior is often disconnected from scales' capabilities and from clinical relevance, and that accurate understanding of weight fluctuation is associated with greater trust in the scale itself. We propose significant changes to how weight data should be presented and discuss broader implications for the design of other ubiquitous health sensing devices.

Author Keywords

Weight, scales, health data perception

ACM Classification Keywords

J.3. Life and medical sciences: Health.

INTRODUCTION

The bathroom scale is the most ubiquitous tool for diagnosing and managing weight issues—arguably, the most ubiquitous health sensor of all—and several studies have shown that frequent weigh-ins help maintain weight loss [25,28]. However, people who are watching their weight often have a marked aversion to stepping on the scale [7]. We hypothesize that some of this resistance comes from the design of the scale's interface. Despite its centrality to global health and wellness, the familiar bathroom scale interface has barely changed since it was first introduced about 100 years ago: it still produces a single value representing one's weight at the moment of measurement. Digital displays

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. UbiComp'13, September 8–12, 2013, Zurich, Switzerland. Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-1770-2/13/09...\$15.00. <http://dx.doi.org/10.1145/2493432.2493456>



Figure 1. Screenshot of a mobile web app used to collect multiple weigh-ins each day. Participants entered their weight and answered three multiple-choice questions at each weigh-in. The result was added to a running graph of weight over time.

have replaced the analog needle, coarse measurements of body fat have been added, and some scales log data for offline review; however, the singular data point is still the main display and is often the only information presented at the time of weigh-in. Most scales answer just one question—“what do I weigh *right now*?”—which may not be the best framing for weight data.

We believe there are several issues with current scales that work against an effective understanding of weight management, which we explore in this paper. For example, digital scale readouts convey an unrealistic level of precision, negatively affecting user perception. We also show that many scale users develop a deep, trusting relationship with their scales despite significant misconceptions about accuracy, trends, and fluctuation; in an online survey of over 800 scale users, we found that respondents with less understanding of how weight fluctuates during the day were less likely to trust their scales. This is exacerbated by the fact that the scale interface makes no attempt to inform users about how weight fluctuates. Our work suggests an opportunity to re-imagine the 100-year-old user interface that is still state-of-the-art in weight management, grounded in best practice in weight management research and consumers' understanding of weight fluctuation. Further, as scales are part of a larger class of increasingly ubiquitous health feedback devices that provide single-point, instantaneous measurements—such as body fat estimators, thermometers, pedometers, and blood pressure cuffs—our work provides a foundation for future design in this broader space.

The rest of this paper is organized as follows: First, we describe related work in weight management (focusing on scales) and intelligibility of ubiquitous interfaces. Second, we analyze a repository of online reviews of scales, examining consumers' understanding of quantitative health measurements in terms of attributes like accuracy, precision, and trends. Third, we outline themes from semi-structured interviews with experts in nutrition on the role of the scale in clinical practice and their clients' relationships with scales. Fourth, we present results from a study quantifying daily weight fluctuation, which has previously been only anecdotally studied even in the clinical literature. Fifth, we describe a large-scale survey of over 800 participants assessing their understanding of how scales operate, how much their weight typically fluctuates, and their own relationships with scales. Finally, we synthesize design recommendations for weight scales and discuss broader implications for the design of health feedback displays.

BACKGROUND AND RELATED WORK

Weight Management and Scales

As links among obesity, mortality, and other health conditions have become clear [2,9], weight management has become a key part of health practice. Obesity is clinically defined in terms of weight and Body Mass Index (BMI) [4,22]; BMI is itself a function of weight and height. Therefore, the scale plays a central role in *diagnosing* obesity. The scale is also used as part of the *treatment* regime for obesity: more frequent use of the scale, such as daily weigh-ins, correlates with better weight maintenance after weight loss [25,28]. Studies have shown people who maintain weight best after weight loss interventions eat healthily, have physical activity in their lives, and regularly monitor their weight [12,27]. Actual approaches to reducing weight are most commonly associated with calorie restriction and increased physical activity [13,16]—i.e., having people eat less food than required to maintain their current weight. Finally, the weight scale also allows a patient or clinician to monitor weight *fluctuation*, which has itself been directly associated with increased mortality [6,21]. Fluctuation is particularly common in individuals dealing with obesity: numerous studies show that successful weight loss is often followed by a recurrence of obesity, with patients sometimes gaining more than they have lost [6,21].

Because caloric restriction seems to have only short-term benefits and often leads to weight regain, and because weight fluctuation is associated with increased mortality, recent work has asked whether weight management should be based more on healthy behaviors than on instantaneous weight [5,18]. In the consumer space, scales such as the Withings and the Fitbit Aria have adopted a self-tracking approach: these scales automatically upload weight and body composition to a website where users can view graphs of their weight over time. However, despite innovations in offline feedback, the fundamental user interface of the scale at weigh-in remains essentially unchanged, reflecting only

instantaneous weight. One exception is a Weight Watchers scale that displays the difference between current weight and a goal weight (or the previous measured weight); however, this still treats single data point measurements as meaningful reflections of current weight and does not inform users of broader patterns of weight fluctuation.

Intelligibility of Feedback in Ubiquitous Computing

One of the core challenges in scale interfaces we will discuss throughout this paper is users' understanding of the underlying data—how weight typically fluctuates and the uncertainty associated with measuring it. Lim and Dey have studied the effects of the *intelligibility* of context-aware systems on user perceptions—essentially, how transparent the reasoning or certainty of these systems are to users [14,15]. They found that exposing the certainty of a system—for example, a confidence region in location-aware systems—improves users' perceptions of the accuracy and appropriateness of a system, so long as the certainty is good enough [15]. In general, the effect of displaying uncertainty on task performance seems to vary by application, sometimes having positive [1] or negative [20] effects.

Other work has looked at using natural-language generation to describe inferences in health data [17,23] as a way to improve human inference. We believe this approach may be promising for weight data, and a systematic understanding of people's grasp of statistical vocabulary is essential to it. Researchers have tried to quantify *words of estimative probability* by having people assign numerical probabilities to words like 'likely', 'uncertain', 'impossible', and so on [11]. Similarly, confusion around measurement descriptions such as 'precision' and 'accuracy' has been explored in science education [24] and in specific scientific domains [26], but we are not aware of similar investigations of lay understanding of such words, despite their frequent use in product descriptions and consumer reviews.

ONLINE REVIEWS STUDY

We began our investigation into users' perceptions of weight scale data with a qualitative analysis of online product reviews from a popular shopping site (amazon.com) for several consumer scales. This study aimed to answer three questions: 1) What are consumers' expectations for accuracy in scales? 2) How do these expectations relate to consumers' satisfaction with devices? and 3) What terminology do consumers use to express these expectations?

We analyzed product reviews for four popular scales: the Withings scale, the Fitbit Aria, a Tanita scale, and a Weight Watchers scale. Amazon.com reviews include two pieces of metadata: a 5-point *product rating* and a yes-or-no *helpfulness rating* (derived from the question "was this review helpful to you?"). The helpfulness rating overestimates the helpfulness of reviews with a small number of positive reviews, so we convert it to a *helpfulness score* by taking the lower bound of its 95% binomial confidence interval.

From a corpus of 1084 reviews, we selected those with at least one helpfulness rating (855 reviews). Of these, we considered only 1-, 2-, 4-, and 5-star reviews (817 reviews) and then coded 100 reviews (the top 50 with 1 or 2 stars and the top 50 with 4 or 5 stars, ordered by helpfulness score). We used affinity diagramming to identify recurrent themes within this subset around users' understanding of precision, accuracy, and uncertainty. We derived a coding scheme from these themes with 44 codes across 5 categories: motivations for using the device, how reviewers test accuracy/reliability, consistency expectations, factors discussed with respect to data quality, and interpretations of noisy data. The reviews were coded, and we used frequency profiling [19] to identify codes that were more frequently found in 4- or 5-star reviews (*positive* reviews) than 1- or 2-star reviews (*negative* reviews), and vice versa.

Results

Trend Focus vs. Data Point Focus

Positive reviews were more likely to exhibit a *trend focus* (28% of positive reviews, 4% of negative reviews). Rather than discussing problems with individual readings, reviewers discussed the overall value of the scale in surfacing fitness trends. For example, from a positive review:

However, body weight fluctuates throughout the day and week. With this scale, I've found myself weighing myself several times per day and looking at my data over a week or month, clear trend lines can be seen despite the daily fluctuations. Ultimately, this is the reason that I bought the scale and makes me very happy.

This reviewer accepts fluctuations in the data, reasoning that the overall trend is more important. In contrast, negative reviews were more likely to quantify the perceived precision of a device and then express a *desire for more consistent readings* (2% of positive reviews, 26% of negative reviews), either within the device or as compared to other devices; for example (from a negative review):

The weight ranges +/- 1.5 lbs each time you use it. So let's say you weight [sic] 150 on the scale at your doctor's office. you can expect your reading to be anywhere between 148.5 to 151.5 when using this scale. [...] I can't rationalize keeping a \$150+ scale that just isn't accurate.

Consumers' expectations for the accuracy and reliability of scales seem to vary depending on their model of use. Those with an understanding of or a focus on trends seem more willing to tolerate noisy data, so long as they can establish a baseline from which to observe change. By contrast, those who gave negative reviews were more likely to focus on perceived noise in the data, even if the magnitude of that noise was similar to that reported in positive reviews.

Vocabulary and Terminology

In total, 68 of the 100 reviews we coded discussed issues around accuracy, precision, or uncertainty. To get a sense of the vocabulary used to express these concepts, we counted

the number of reviews containing various words and their derivatives (we list words here only by one form, e.g. consistency for consistent/consistency and derivatives). By far the most-used term was 'accuracy' (in 48/68 reviews), followed by 'consistency' (22/68), 'fluctuation' (10/68), 'variance' (8/68), 'precision' (6/68), 'reliable' (5/68), and 'repeatable' (3/68). We note that even in this small sample, words were not used consistently by reviewers: for example, 'precision' was used to refer both to the concept of accuracy and of precision by different reviewers. We also observed a strong preference for the use of the term 'accuracy' to refer broadly to issues of measurement uncertainty. We therefore believe that a more systematic investigation of vocabulary for expressing uncertainty is warranted.

EXPERT INTERVIEWS

We interviewed four experts on weight change to validate the findings from our online review study, to better understand how scales are used in weight management, and to learn how experts see the effects of scale use on their users:

- E1, a professional strength and nutrition coach, works with clients trying to lose weight and clients trying to add muscle mass for specific athletic activities.
- E2, a dietician whose practice includes both athletes and non-athletes dealing with body weight issues. She is also an author of two cookbooks on healthy eating.
- E3, an osteopathic physician who works in a family medical practice and focuses on weight loss issues. He works in a low-income area with high rates of obesity.
- E4, the author of popular books and a blog on nutrition practices and a practicing fitness and nutrition coach. He primarily works with clients looking to lose weight.

We conducted a semi-structured interview with each expert, focusing on their background, perceptions of scales, how scales fit into their practice, and their clients' perceptions of weight and scales. We used affinity diagramming of transcripts to identify high-level themes, discussed below.

Results

Scales Can Reinforce Inappropriate Goals.

E1 and E2 both stressed that while weight is important, it is not always a complete picture of clients' progress toward fitness goals. E1 noted that many people do not make the connection that body composition is often more important than weight and that "there's people that completely change their body composition and stay the same weight." E2 also noted that people use weight as an "inappropriate goal". One of her clients was "hung up" because she couldn't get to 125 lbs, even though in photos she clearly had a lean body composition. E2 stated that a specific weight—as a number—is often "such an identity for people", and that people are "not so obsessed with your shoe size". E4 called these "assumed" numbers: "a lot of people decide on a number at the beginning that they think they will look good

at". These issues were reflected in how E1, E2, and E4 use weight with clients: as one measure amongst several, including body fat calipers (E1 and E2) and circumference measures (E1, E2, and E4), e.g. waist or shoulder circumference. E4 noted, "weight is an excellent tool when used in combination with other metrics".

Emotional Connection

E2's observation that weight can act as an "identity" for people reflects a broader theme of emotional connections to scales and weight that pervaded our discussions with experts. E1, E2, and E3 discussed how they must tailor their recommendations to clients, depending on how comfortable they estimate each client will be with regular weighing. E1 noted that weighing daily would drive most people "batty"; "they have an emotional experience... they see numbers and it's not what they expect"; and that weight can move "wildly" for some clients; e.g., simply by changing the proportion of carbohydrates in one's diet, a person might see a change of 5–8 lbs. E1 described one client:

There was a fellow that was ignoring the other measures [he only looked at weight]... He was trying to lose weight, and he gained a pound. He was blaming external forces, he was venting: "This isn't working!"... I pointed out, "Well, you lost a few inches off your waistline." It was a very emotional reaction from a level-headed guy.

Overreaction to Fluctuations

E2 noted that people react "out of proportion" to small changes in weight of 1–2 lbs and they "extrapolate forward in their minds". She described clients as getting "the horrors" when they feel like their weight moves in an undesirable direction. E4 noted people can get "kind of crazy" and tend to think of small weight changes as absolute instead of transient. He has to tell them: "let's wait a day or two and see what happened". He also noted a tendency for some people to weigh themselves at home and the gym and worry about differences of a pound or two without considering differences in the scales used. E1 and E3 both tailored their recommendations to their estimation of a patient's ability to handle regular weighing; as E3 noted: "some people get bent out of shape if they weigh themselves every day".

Regular Weighing Still Has Significant Value

Despite the potential issues with weighing our experts outlined, all of them considered it an important practice and recommended most clients weigh themselves about once a week. Recognizing the tendency for weight to fluctuate during the day from their own experience, they suggest clients weigh in at a consistent time of day and under similar conditions (e.g., just before breakfast) and typically once a week (E1 estimated daily fluctuations at 3–5 lbs, and E4 at 3–4 lbs, though neither were aware of studies measuring this fluctuation). At the same time, E1 noted the potential value of weighing more often: "if they can mentally take it, I tell them to go every day: you can see amazing trends." He even described some clients who weigh multiple times a

day: "They really start to connect to how certain behaviors and food choices affect data", but noted that while some people get excited by connecting data to behaviors or conducting self-experiments, there is a personality split: this sort of tracking works more for people who have "a bias towards data", a split also noted by our other experts.

Finally, E4 stated, "the place where I like it [the scale] is, after getting to a good point, understanding what a healthy weight range is." He described scales as particularly valuable for supporting weight maintenance among people who have lost weight: once people get to a steady weight and establish a healthy weight range, they can see when weight gets to "an amount outside of a comfortable zone" then adjust their behavior. In general, our experts cast the best use of weight as an indicator of a trend rather than as an absolute value; as E4 said: "We only really want to know: would that line be 'kind of going down' or 'kind of going up'".

Education and Rationale are Essential

E2 and E3 both emphasized the importance of educating clients to help them understand weight changes. E3 noted that "a third to a half of a visit" typically consists of providing background information—for example, if a client gains a couple of pounds, E3 has to explain that it is probably water. E2 echoed this sentiment when talking about client compliance: "Mandates don't work. When you explain why, you get better compliance". All experts discussed the need to explain potential sources of weight fluctuation to clients as a way to allay their concerns about small changes in weight. These practices suggest that perhaps approaches to conveying intelligibility—particularly rationales or explanations of *why* data looks as it does [14]—may have strong impact in the weight space.

WEIGHT TRACKING STUDY

The results of the online reviews study and our expert interviews support our hypothesis that a significant number of consumers have misperceptions about scale accuracy and weight fluctuation. However, we cannot accurately assess people's understanding of daily weight fluctuation without some standard against which to judge their perceptions. We were unable to find studies of *within*-day weight fluctuation in the literature (weight change is typically studied between days). Furthermore, consultations with physicians and dietitians suggested such data could help them allay clients' concerns, but they were not aware of any studies that had collected it. To begin to fill this gap in the literature, we devised a study to gather data on within-day weight fluctuation. We specifically sought to answer two questions: 1) How much does a person's weight typically vary during a single day? and 2) How much do weighing conditions like clothes or the scale used affect weight measurements? Both of these questions inform our hypotheses that single-point, context-free measurements overlook important aspects of weight management and that consumers place undue emphasis on numerical precision in weight measurements.

Component	Effect (lbs)	SD		
Clothing			$F_{2,641} = 31.32$	$p < .0001$
partially	0.85	0.30	$t_{641} = 2.81$	$p < .01$
fully	2.17	0.28	$t_{641} = 7.71$	$p < .0001$

Table 1. Effects of weighing conditions on weight.

We used a journaling approach to collect multiple weigh-ins from users on a mobile web app (Figure 1). We recruited within our institution (via a departmental email list) and on weight-related Internet forums. For participants within our institution, we placed 10 digital weight scales of the same model throughout our building in easily accessible areas: kitchenettes, locker rooms, and the building foyer. Participants were not compensated but were presented with graphs of their own data as an enticement for the curious (Figure 1). We asked participants to weigh themselves at least 3 times daily for a period of at least 10 days, spanning two weekends, and to use our web app to report their weight immediately after weighing. In addition to the user’s current weight, our phone app requested *clothing state* (“fully”, “partially”, or “not”), *scale* (“work”, “home”, or “other”), and *phone presence during weighing* (“present” or “not present”). Time of entry was logged automatically.

After excluding participants that provided three or fewer readings, we had data from 23 participants (69% male): 17 internal to our organization and 6 external. Participants weighed themselves an average of 28.8 times (sd=23.8, min=6, max=109); 15 participants gave us at least 20 measurements. Mean weight among participants was 168.2 lbs (sd=8.5), mean age was 32.5 (sd=9.4).

Results

Effects of Weigh-in Conditions

Understanding the effects of weigh-in conditions (clothes, scale, etc.) would allow us to better explain potential causes of weight fluctuation to users. We used a mixed-model regression and analysis of variance to analyze the effects of *clothing* and *scale* on *weight*. *Clothing* was modeled as a fixed effect, allowing us to estimate the average effect of wearing clothes across all participants. *Participant* and *scale* (nested within *participant*) were modeled as random effects, allowing our model to account for the effect of each person’s scale separately. Before running this model, the effect of *phone presence* was accounted for by subtracting the mean weight of a smartphone—0.29 lbs (sd=0.05)—taken from a database (<http://smartphones.findthebest.com>) of 464 models of smartphone. The effects of model components are summarized in Table 1: on average, being *partially* clothed increased weight by 0.85 lbs and being *fully* clothed increased weight by 2.17 lbs. Our model also estimates an offset for each scale from the correct weight. The offset range was 4.56 lbs (IQR=1.33 lbs). This is fairly consistent with previous work that found digital scales in a hospital had a range around the standard weight of 5.51 lbs (IQR=1.15) [8], supporting our model’s validity.

Within-day range	Mean (lbs)	SD	Min	Max
unadjusted	3.60	2.22	0.40	11.00
adjusted	2.72	1.88	0.40	11.87

Table 2. Unadjusted and adjusted within-day weight ranges.

Within-Day Weight Variation

To estimate typical within-day weight variation, we considered all instances of any participant submitting at least 3 weigh-ins in a calendar day. We then calculated the difference between the maximum and minimum recorded weight for each day; we call this the *within-day range*. Our model of clothing and scale effects also allows us to also derive an *adjusted weight* for each weight. We do this by subtracting the effect of the participant’s recorded clothing level, scale used, and phone presence from each weight. Using these *adjusted weights*, we can calculate an *adjusted within-day range*. While this adjusted range should more closely approximate actual weight fluctuation, the unadjusted range reflects what a scale user is more likely to observe in practice. Therefore, we report both (Table 2, Figure 2a). The mean within-day range was 3.60 lbs (2.72 lbs adjusted), validating our experts’ estimates of about 3–5 lbs.

These results suggest body weight can fluctuate substantially throughout the day. On top of that, changing clothes or weighing on a different scale may have a significant effect on the weight shown on a scale, even if body weight has not changed. Given that product reviews from our first study suggest even changes of a single pound may be important to users, these results indicate that daily observed weight variation could cause undue concern amongst people who weigh themselves often (or with different scales) but who do not fully understand these sources of weight change.

Weight Range by Mean Weight

We also hypothesized that heavier individuals might see a greater within-day weight fluctuation, implying that it would be better to examine within-day weight fluctuation as a percentage of each individual’s mean weight. Somewhat to our surprise, we found no evidence of a correlation between an individual’s mean weight and their mean within-day weight range ($F_{1,19}=0.0001$, $R^2=-0.05$, $p=0.99$). While we saw no evidence for such a relationship, we note that we had no participants with a mean weight over 300 lbs. It is possible that in those with very high (or low) weight, fluctuation patterns differ from those observed in our sample.

Focus & Limitations

We stress that the regression used in the study was only to approximate the fluctuation in weight measurements, as our primary focus is on examining the appropriateness of instantaneous measurements of weight from an end-user perspective. That is, the physiological influences on weight fluctuation (menstrual cycle, salt intake, etc.) are not in our scope: we wanted to know what people’s weight fluctuations look like to them, regardless of what caused them. Our

focus only on fluctuation—not on causes of fluctuation—precisely complements our observation that scales do not use or present any of this contextual information either.

SCALE PERCEPTIONS SURVEY

A pervasive theme throughout our investigation was users' struggle to understand and account for fluctuations in data: both in product reviews and expert interviews, we encountered mismatches between the magnitude of reactions to weight change and the actual significance of that change, given our knowledge of weight fluctuation derived from our weight tracking study. We conducted an online survey to better gauge the relationship between scale users' perceptions of weight data and their understanding of weight fluctuations—e.g., do people with a better understanding of weight fluctuation trust their scales more? Noting the inconsistent use of statistical vocabulary by product reviewers, we were also interested in establishing a common lay vocabulary for scale properties like accuracy and reliability.

We recruited via mailing lists within our organization, on weight- and fitness-related forums, and on Twitter. Internal participants were offered a \$10 gift card; external participants were entered into a raffle for a \$50 gift card. We also invited participants in our weight tracking study to complete an exit survey that included about that study as well as all questions from the scale perceptions survey. These participants were offered the same compensation as survey-only participants for completing the exit survey.

Results

Of 892 total respondents, 18 had been participants in our weight tracking study and 30 were internal to our institution. Of the 861 others, 716 were recruited via E4, who advertised our survey to his mailing list. 59% were male and 79% weighed themselves regularly. 67% reported they were trying to lose weight, 15% to maintain weight, 5% to gain weight, and 9% had other goals (e.g., changing fat/muscle composition). The next three subsections address respondents' understanding of weight fluctuation, the connection between that understanding and their perceptions of scales, and common vocabulary for scale accuracy and reliability.

Understanding of Within-Day Weight Fluctuation

To estimate respondents' understanding of typical daily weight fluctuation, we prompted them with the following:

Imagine your heaviest weight on a typical day and your lightest weight on the same day. Please indicate how likely you think each of the following scenarios is.

Respondents then indicated whether they thought each of the following scenarios was *very likely*, *somewhat likely*, *somewhat unlikely*, or *very unlikely*:

- *Your heaviest weight is more than 10 lbs (4.5 kg) higher than your lightest weight.*
- *Your heaviest weight is 8 lbs (3.6 kg) higher than your lightest weight.* (This question was repeated for 6 lbs, 4 lbs, and 2 lbs.)

- *Your heaviest and your lightest weight are the same.*

In essence, we wanted respondents to indicate their expected distribution of within-day weight ranges. Results of these questions are shown in Figure 2b alongside the distribution of within-day ranges from our weight tracking study.

Respondents' estimations of within-day weight range were generally good: the shape of their average estimated distribution is similar to our observed distribution. Respondents tended to place 2 lbs or 4 lbs as the most likely weight range, close to our observed 3.6 lbs (2.72 lbs adjusted). However, many still over-estimated both the chances that no weight difference would be observed or the chance that a much larger difference (e.g. 8 or 10 lbs) would be observed.

Weight Fluctuation Knowledge and Weight Data Perception

To compare responses between respondents who had a more or less accurate understanding of daily weight fluctuation, we categorized their likeliness estimates into *accurate* and *inaccurate estimates*. *Accurate estimates* were those that: (1) rated the 0, 8, and 10 lbs ranges as *Very* or *Somewhat unlikely*, and (2) rated the 2 and 4 lbs ranges as *Very* or *Somewhat likely*. We did not factor the 6 lbs range into our categorization. Given this categorization, 326 respondents (36.5%) had *inaccurate estimates* and 566 (63.5%) had *accurate estimates*, suggesting that a majority had a good understanding of typical within-day weight fluctuation. While this may not be surprising in a population where most people weigh regularly (and we do not claim that this generalizes broadly), it is noteworthy that even in this population 36.5% of people had *inaccurate estimates* of weight fluctuation. To investigate the effect of this knowledge on perceptions of weight data, we asked respondents four Likert-scale questions on their attitudes toward scales: *unreliability*, *trust*, *worry*, and *eagerness* (Figure 3).

To analyze the Likert data, we used the Aligned-Rank Transform (ART), which allows nonparametric testing of multiple factors with an ANOVA. We included *range estimate quality* (*accurate* or *inaccurate*) and *weighs regularly* (*yes* or *no*, self-reported: "Do you weigh yourself regularly (for example, once a week or more)?") and their interaction

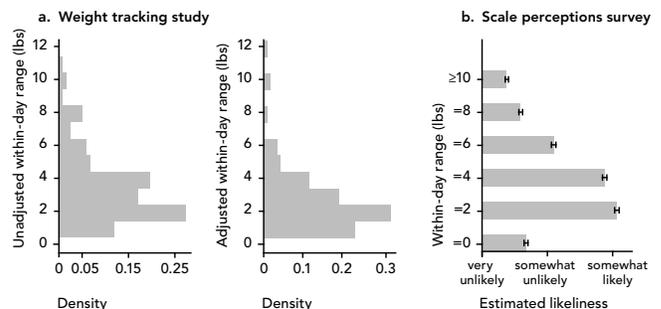


Figure 2. a) From the weight tracking study: histograms of within-day weight ranges (max - min weight within a day) before and after adjustment for weigh-in conditions. b) For comparison, from the scale perceptions survey: respondents' estimated likeliness of various within-day weight ranges.

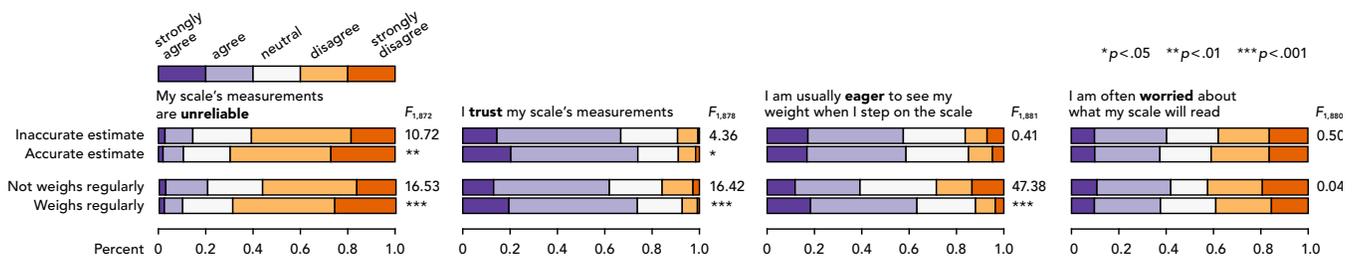


Figure 3. Results of the four Likert-scale questions on scale attitudes, broken down by the quality of the respondents' estimation of within-day weight fluctuation and by whether or not respondents weighed themselves regularly.

as factors in our analysis. We included the latter factor because regular weigh-ins improve weight change outcomes [25,28], and we were curious if it was associated with perceptions of weight data. We did not find a significant *range estimate quality* × *weighs regularly* interaction effect on *unreliability* ($F_{1,872}=.31, n.s.$), *trust* ($F_{1,878}=.04, n.s.$), *worried* ($F_{1,880}=.38, n.s.$), or *eagerness* ($F_{1,881}=2.80, n.s.$). Both having an *accurate range estimate* and *weighing regularly* significantly decreased *unreliability* and significantly increased *trust*. *Weighing regularly* also significantly increased *eagerness* (Figure 3).

We also included *gender* (male or female) and recruitment origin (via E4 or all others) in the model. Gender had a significant effect on *eager* and *worried*: women were less eager to step on the scale and more worried at what it would read, consistent with previous literature [7]. *Origin* had a significant effect on *worried* (Figure 4): those recruited via E4 were more worried at what the scale would read, but had no other significant differences from other respondents.

We also asked the first two Likert questions (*unreliability* and *trust*) to participants in our weight tracking study in a pre-study survey (Figure 5). Because these participants also filled out our weight perceptions survey as an exit survey, we were able to see whether tracking their weight and seeing their graph of daily fluctuation had an effect on their attitudes towards scales. Paired Wilcoxon signed rank tests found a significant decrease in *unreliability* ($V=5, p<.05$) from the *pre-* to *post-study* surveys, but no significant difference in *trust* between *pre-* and *post-study* surveys.

Common Vocabulary

We sought to find a common vocabulary to communicate concepts like the accuracy or precision of a scale to consumers. In particular, we were interested in scale accuracy, measurement reliability (both internal and external), and scale readability (also called resolution). We created four definitions to reflect these concepts as applied to scales and refined them through survey piloting (Table 3).

We asked respondents a series of multiple choice questions in the form “The *precision* of a scale refers to...”, where “precision” was replaced with one of five words and the respondent selected one of our four definitions. Respondents were instructed that “This is not a test: we are interested in how you think about these words.” Respondents were asked to define *precision*, *accuracy*, *consistency*, *resolution*,

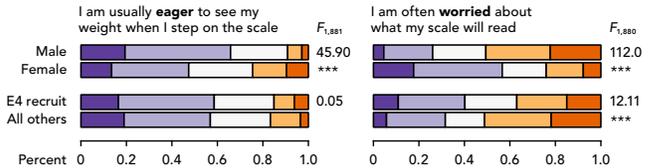


Figure 4. Gender and recruitment strategy differences.

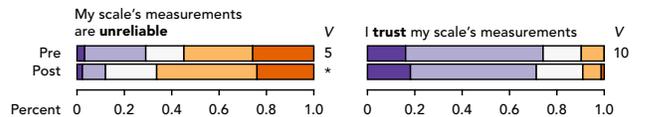


Figure 5. Likert questions on scale attitudes asked before and after participation in the weight tracking study (N=15).

and *repeatability*. All of these words (or derivatives) appeared in a subset of the product reviews we examined and were used by reviewers to refer to some property of the data, although not necessarily in a consistent manner. Table 3 shows the results of the vocabulary questions, with the most common definition for each word highlighted.

Discussion

We found that those who weigh in regularly trust their scales more, are less likely to believe that scales are unreliable, and generally report more eagerness to step on the scale. This is not surprising, and it is difficult to assign causality here, as it seems obvious that someone who is eager to step on the scale would also weigh more often. However, it is important to note these correlations, as they are consistent with previous work suggesting better outcomes for those who weigh in more often [25,28].

Nearly 40% of our population did not have accurate estimates of within-day weight range. We found that those who did have accurate estimates of daily weight fluctuation also believed their scale's measurements were more reliable and had greater trust in those measurements and that this effect was independent of whether or not those people weighed regularly. While it is also difficult to assign causality here, when considered alongside our other results, this is suggestive: in our product review study, we found reviewers were concerned about fluctuations in weight that our tracking study suggests are typical, and our experts find it valuable to educate people about daily weight fluctuations to increase compliance and allay concerns. In this study, we found that people with greater understanding of those fluctu-

Definition	accuracy	precision	consistency	repeatability	resolution
1. how close a single measurement of an object on that scale is to that object's actual weight	73.7%	26.3%	6.1%	7.7%	11.5%
2. how close several measurements of the same object on that scale are to each other	6.8%	15.7%	75.3%	75.0%	9.2%
3. how close a measurement of an object on that scale is to measurements of the same object taken on other scales	11.8%	11.9%	14.7%	12.4%	20.3%
4. the smallest change in weight that can be detected using that scale	7.6%	46.0%	3.9%	4.8%	59.0%

Table 3. Proportion of respondents assigning each definition to each word. The most popular definition for each word is in bold.

tuations express greater trust in scale data and find it more reliable. This suggests their more sophisticated understanding of the underlying data makes them more able to trust the data they see, even when it fluctuates.

DESIGN RECOMMENDATIONS

In this section, we synthesize design recommendations for weight scales based on all of our study results. We believe with further research many of these recommendations could be generalized to other single-point health sensors as well.

Vocabulary Recommendations

The following vocabulary recommendations are based on the majority definition assigned to each word:

- Use **accuracy** as it is typically used in statistics; that is, to refer to how close a measurement is to its actual value.
- Use **consistency** in place of ‘precision’ or ‘repeatability’; that is, to refer to how close repeated measurements on the same device are to each other. While ‘repeatability’ offered similar agreement with this definition, ‘consistency’ was more often used in product reviews and therefore we believe is a more widely understood term. Interestingly, in statistics, ‘precision’ is often given this definition [3], but was only selected by 15.7% of respondents.
- Use **resolution** to refer to the smallest change that can be detected with a measurement device (with some caution).
- Do not use **precision**, as it is too often confused with several different concepts. We note that confusion between ‘precision’ and ‘accuracy’ has been recognized in other domains [24]; our results suggest that the confusion may be primarily one-way: ‘precision’ is often used to mean ‘accuracy’, but the opposite is not true.

Unfortunately, our results do not suggest a strong candidate to be used to express the concept of reproducibility (definition #3). We consider these recommendations an important starting point for exploring natural language feedback techniques in weight scale design, including devices that take a more educational approach to weight data.

Reflect Data Uncertainty

Our results suggest that traditional scales do not adequately convey uncertainty: as found in online product reviewers’ concerns about accuracy—suggesting scales are not conveying accuracy well—and in experts’ discussion of emotional reactions to weight fluctuation—“the horrors”. Reinforcing these results, Our online survey found that greater

knowledge of weight fluctuation was associated with higher trust of scales. Here we offer specific recommendations to improve how scales reflect uncertainty based on our results.

Avoid false precision in single-point measurements: Digital scales, even inexpensive ones, typically have quite fine resolutions (0.2 lbs or less). However, our weight fluctuation study suggests that reporting instantaneous weight down to 0.2 lbs gives a false sense of precision: body weight typically fluctuates by 3 or 4 lbs on a given day, and most people weigh at most once a day. We recommend against reporting weight to fractions of a pound and suggest instead reporting at a resolution more appropriate for daily measurement (perhaps 1 or 5 lbs) or using ranges instead of point measurements. This false precision at the moment of weigh-in reinforces the harmful “weight as identity” paradigm noted by our experts, in which people identify with a particular weight they want to be rather than focusing on healthy change. We do not believe it is sufficient to address these issues through supplemental user interfaces (e.g. as Withings or Fitbit Aria scales do with web-based graphs of data), but that false precision and a focus on single point measurements must be addressed *at the moment of weigh-in*.

Adopt an explicit model of weighing frequency: Our experts stressed the importance of tailoring weigh-in frequency to users and often talked about the daily or weekly weigh-in model. Given how common this model is, and how uncommon multiple daily weigh-ins are, these results further stress a movement away from instantaneous measurement. A scale that adopts an explicit model of the frequency of weigh-ins (even with as simple a cue as “your weight today:” or “your weight this week:”) would reinforce experts’ models. At the same time, such a model could more effectively avoid the problems of false precision in two ways:

- With an explicit model of weighing frequency, empirical data on weight fluctuation (such as that from our weight tracking study, or from the user themselves) could be used to set the precision reflected by the scale.
- Multiple weigh-ins within a single period (e.g., same day or week) could be used to generate a more accurate average measurement rather than separately reporting instantaneous measurements. This avoids issues of stress over accuracy observed in our product reviewers and reported by experts and explicitly enforces data use patterns advocated by our experts.

Such a model could be used to shift the focus from single data points to trends, which are more suited for weight maintenance and the idea of staying within a desired range.

Educate users about uncertainty: We saw in our online vey that users with a greater knowledge of within-day weight fluctuation had greater trust in their scales' accuracy. Similarly, in our weight tracking study, participants' perceptions of the unreliability of their scales decreased after seeing regular graphs of their daily weight fluctuation. Our experts also stressed the importance of education to adoption of healthy patterns of scale use. We therefore suggest not only addressing problems of false precision and lack of an explicit model of weighing frequency as described above, but also to use the opportunity of a regular weigh-in to educate users about their weight, contextualizing how much their weight typically fluctuates, and exploring possible causes of this fluctuation. This may be as simple as textual explanations: "We estimate your daily weight to within 3 lbs since weight typically fluctuates about that much during the day", or may involve graphical depictions of weight variability. Indeed, our lightweight graphing approach in the weight tracking study could be considered a rough first pass at educating users about their weight fluctuation.

There has been extensive work in visualizing uncertainty; see [29] for an overview of several techniques. Some of this work (e.g. the use of shading to visualize probability distributions of point estimates or continuously varying uncertainty [10]) may be particularly applicable to the weight domain and to implementing the above recommendations.

DESIGN VIGNETTES

We also developed several design vignettes to illustrate the potential value of our design recommendations and to initiate future thinking about design in this space. We imagine:

- A scale that shows a moving average of weight in place of the current weight and asks for more readings if the bounds of the estimate are large. Such a scale might only give a range (visually or numerically) and *never provide point estimates of current weight (instantaneous or averaged)*. We encourage designers of scales to reflect data uncertainty clearly, and we consider the lack of this in current scale designs to be a critical issue. This vignette aims to *avoid false precision in measurements*.
- An "always-on" scale that provides a current estimate (with uncertainty) of the user's weight—whether they have stepped on it recently or not—updated based on a model of how the person's weight has changed in the past. Stepping on the scale at any time would update the model and the current estimate. The model could pull in additional information through occasional, low-burden questions, like "how well did you sleep last night?", "when did you eat yet today?", or "have you changed what you are wearing?"; it could also educate users about weight fluctuation: "You are weighing in the afternoon,

but last weighed in the morning. Had you weighed this morning, we estimate your weight would be X ".

Such a scale illustrates the power of all of our recommendations: it *avoids false precision* through estimates with uncertainty, but also *adopts a model of weighing frequency* and begins to *educate users about uncertainty* both implicitly (through the increase in the magnitude of the uncertainty with less frequent weigh-ins) and explicitly (through questions and textual feedback).

- A digital scale that graphically emulates a traditional analog scale, with exaggerated needle movement and width to reflect uncertainty. For example, the needle might oscillate in proportion to uncertainty or continuously oscillate across a predicted range rather than settling on a single value. Such a scale might be a more natural way to represent variability in weight data.

The first vignette in particular—omitting the ubiquitous "current weight"—represents a radical departure in design that we believe is supported by our findings and will stimulate discussion and design exploration in this space.

DISCUSSION & CONCLUSION

In this work, we considered the impact of the scale's user interface on weight perceptions through four complementary studies, making several contributions to redesigning the interaction of the ubiquitous bathroom scale. First, we contribute several findings that demonstrate the impact of users' models of accuracy and weight fluctuation on their perceptions of scale reliability, and show that gaps in this understanding cause disproportionate emphasis on single readings and frequently lead to understandable discouragement with small weight fluctuations. Second, we contribute three design recommendations to address the shortcomings of scale feedback that exacerbate these gaps in users' understanding: avoid false precision, adopt an explicit model of weighing frequency, and educate users about fluctuation; and we demonstrate how these guidelines might be adopted through design vignettes. Finally, we provide vocabulary recommendations for describing scale properties, crucial to the increasing interest in natural language health feedback.

Our findings suggest a better scale can be designed without investing in more expensive equipment, better calibration, or even clearer instructions for obtaining better data (e.g. to always use the scale on a hard surface), but instead by handling the resulting data and consequent user feedback in a more considered way. Scales are already ubiquitous, cheap, and fairly accurate; greater gains may be had by pushing the state of the art in feedback. We encourage designers to think about how to design better scales given their inherent uncertainty and the properties of the data being collected.

This work may also provide a foundation for future work exploring the design of other ubiquitous health sensors that employ instantaneous feedback, such as thermometers, blood pressure cuffs, and blood glucose monitors. We expect that similar issues—such as gaps in knowledge of the

accuracy of a device or what constitutes typical fluctuation—may manifest in these domains. Further work is necessary to understand how our design recommendations generalize to these systems, but better reflecting the underlying model and educating users about uncertainty will likely remain crucial. As ubiquitous computing continues to push low-cost novel health sensing into the wild, it is important to consider how the accuracy of these systems affects users, and how much we can gain from more intelligent feedback.

Finally, as our experts noted, a particular problem with using weight as a proxy for other goals—change in appearance, cardiovascular risk, etc.—is its approximate relationship with these outcomes. The ubicomp community has an opportunity to revisit the underlying health questions users want answered, what data we can use to answer them, and how to help interpret that data—rather than falling back to the nearest convenient measure (e.g., weight) and simply reporting individual data points without context or interpretation. Instead, we can use more appropriate sensors—with more intelligent feedback—to give better, clearer answers to a person’s underlying health and wellness concerns.

REFERENCES

- Antifakos, S., Schwaninger, A., and Schiele, B. Evaluating the Effects of Displaying Uncertainty in Context-Aware Applications. *Ubicomp 2004*.
- Barness, L.A., Opitz, J.M., Gilbert-Barness, E. Obesity: genetic, molecular, and environmental aspects. *Am J Med Genet A*. 2007 Dec 15;143A(24):3016–34.
- BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML. The international vocabulary of metrology—basic and general concepts and associated terms (VIM3). JCGM 200:2012.
- CDC Guidelines for Defining “Overweight” and “Obesity”. <http://www.cdc.gov/obesity/adult/defining.html> .
- Clark, M. Is weight loss a realistic goal of treatment in type 2 diabetes? The implications of restraint theory. *Patient Education and Counseling* 53, 3 (2004), 277–83.
- Diaz, V.A., Mainous, A.G., and Everett, C.J. The association between weight fluctuation and mortality: results from a population-based cohort study. *J Community Health* 30, 3 (2005), 153–65.
- Drury, C.A.A. and Louis, M. Exploring the association between body weight, stigma of obesity, and health care avoidance. *J Am Acad Nurse Prac* 14, 12 (2002), 554–61.
- Goldberg, R. and Hebbard, G. How accurate are hospital scales? *Med J of Australia* 194, 12 (2011), 665.
- Haslam, D.W., James W.P. (2005). Obesity. *Lancet* 366 (9492): 1197–209.
- Jackson, C.H. Displaying Uncertainty With Shading. *The American Statistician* 62, 4 (2008).
- Kent, S. Words of Estimative Probability. *Studies in Intelligence*, (1964).
- Kimokoti, R.W., Newby, P.K., Gona, P., *et al.* Diet quality, physical activity, smoking status, and weight fluctuation are associated with weight change in women and men. *Journal of Nutrition* 140, 7 (2010), 1287–93.
- Kruger, J., Galuska, D.A., Serdula, M.K., Jones, D.A. Attempting to lose weight: specific practices among U.S. adults. *Am J Preventative Med* 26, 5 (2004), 402–6.
- Lim, B.Y. and Dey, A.K. Assessing Demand for Intelligibility in Context-Aware Applications. *Ubicomp 2009*.
- Lim, B.Y. and Dey, A.K. Investigating Intelligibility for Uncertain Context-Aware Applications. *Ubicomp 2011*.
- Nicklas, J.M., Huskey, K.W., Davis, R.B., and Wee, C.C. Successful weight loss among obese U.S. adults. *Am J Preventative Med* 42, 5 (2012), 481–5.
- Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., and Sykes, C. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artif. Intell.* 173, 7–8 (May. 2009), 789–816.
- Provencher, V., Bégin, C., Tremblay, A., Mongeau, L., Boivin, S., and Lemieux, S. Short-term effects of a “health-at-every-size” approach on eating behaviors and appetite ratings. *Obesity* 15, 4 (2007), 957–66.
- Rayson, P. and Garside, R. Comparing Corpora using Frequency Profiling. *CompareCorpora* 2000, 1–6.
- Rukzio, E., Hamard, J., Noda, C., and Luca, A. De. Visualization of Uncertainty in Context Aware Mobile Applications. *MobileHCI 2006*, 247–250.
- Rzehak, P., Meisinger, C., Woelke, G., Brasche, S., Strube, G., and Heinrich, J. Weight change, weight cycling and mortality in the ERFORT Male Cohort Study. *European J Epidemiology* 22, 10 (2007), 665–73.
- Sweeting, H.N. Measurement and Definitions of Obesity In Childhood and Adolescence: A field guide for the uninitiated. *Nutrition Journal* 6, 32 (2007).
- Tollmar, K., Bentley, F., and Viedma, C. Mobile Health Mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device. *Pervasive Health 2012*.
- Tomlinson, J., Dyson, P. & Garratt, J. (2001). Student misconceptions of the language of error. *U Chem Ed* 5, 1–8.
- Vanwormer, J.J., Linde, J.A., Harnack, L.J., Stovitz, S.D., and Jeffery, R.W. Self-Weighing Frequency Is Associated with Weight Gain Prevention over 2 Years Among Working Adults. *Intl J Behavioral Med*, (2011).
- Walther, B.A., and Moore, J.L. The concepts of bias, precision, and accuracy, and their use in testing the performance of species richness estimators. *Ecography* 28 (2005).
- Wang, X., Lyles, M.F., You, T., Berry, M.J., Rejeski, W.J., Nicklas, B.J. Weight regain is related to decreases in physical activity during weight loss. *Medicine and science in sports and exercise* 40, 10 (2008), 1781–8.
- Wing, R.R., Tate, D.F., Gorin, A.A., Raynor, H.A., Fava, J.L., Machan, J. STOP regain: are there negative effects of daily weighing? *J Consulting and Clin Psych* 75, 4 (2007).
- Zuk, T. and Carpendale, S. Theoretical Analysis of Uncertainty Visualizations. *SPIE-IS&T Electronic Imaging* 2006.