Designing for User-facing Uncertainty in Everyday Sensing and Prediction

Matthew Jeremy Shaver Kay

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Julie A. Kientz, Chair

Shwetak N. Patel, Chair

James Fogarty

Program Authorized to Offer Degree:

Computer Science & Engineering

University of Washington

Abstract

Designing for User-facing Uncertainty in Everyday Sensing and Prediction

Matthew Jeremy Shaver Kay

Chairs of Supervisory Committee:

Associate Professor Julie A. Kientz
Human Centered Design & Engineering

WRF Entrepreneurship Endowed Professor Shwetak N. Patel
Computer Science & Engineering

As we reach the boundaries of sensing systems, we are increasingly building and deploying ubiquitous computing solutions that rely heavily on inference. This is a natural trend given that sensors have physical limitations in what they can actually sense. Often, there is also a strong desire for simple sensors to reduce cost and deployment burden. Examples include using low-cost accelerometers to track step count or sleep quality (Fitbit), using microphones for cough tracking [57] or fall detection [72], and using electrical noise and water pressure monitoring to track appliances' water and electricity use [34]. A common thread runs across these systems: they rely on inference, hence their output has uncertainty—it has an associated error that researchers attempt to minimize—and this uncertainty is user-facing—it directly affects the quality of the user experience. As we push more of these sorts of sensing and prediction systems into our everyday day lives, we should ask: does the uncertainty in their output affect how people use and trust these systems? If so, how can we design them to be better?

While existing literature suggests communicating uncertainty may affect trust, little evidence of this effect in real, deployed systems has been available. I wanted to start with the simplest, most ubiquitous sensing system I could think of to see if a failure to communicate uncertainty might affect trust or usage. I turned to the home weight scale, a sensing system that has been with us for at least 100 years, but which maintains at the moment of measurement about the simplest feedback interface possible: it just tells you your weight! I conducted a series of studies, starting with qualitative investigations into people's understanding of uncertainty in weight scales, through a study of actual variability in scale measurements, to an online survey of attitudes towards scales from a large number of scale users. I found that many people judge their scales by their perceived uncertainty, but often confuse aspects of error like bias and variance when doing so. I found that people who have a better understanding of weight variability trust their scales more. I found that the scale does little to help people along in their understanding of error at the moment of weigh-in. The design of the scale is today causing problems with trust and abandonment. How can we design it better?

To investigate how to communicate uncertainty effectively to people, I turned to another domain: realtime transit prediction. I wanted to understand how to communicate uncertainty in a way that people grasp more viscerally. I sought to develop a visualization of a continuous measure (time to arrival) that capitalizes on a frequency-based (or discrete) representation to improve people's probabilistic estimates. I introduced a novel visualization technique, quantile dotplots, that capitalizes on both a discrete presentation of uncertainty and people's fast-counting abilities (called subitizing) to improve the precision of their probabilistic estimates by about 15%. If my work in weight scales asks what happens when we get it wrong, this work aims to build an understanding of how to get it right.

Finally, effective design for user-facing uncertainty is not limited to the visual communication of uncertainty: it is not enough to slap an effective visualization on top of whatever model exists in a predictive system. In even conceptually simple prediction tasks—*is it going to rain today?*—people exhibit preferences for different types of errors. In precipitation prediction, this manifests as *wet bias* [82]—the tendency of weather forecasters to over-predict the probability of rain to reduce the chance that people do not prepare for rain and then blame the forecaster when they are rained on. Even an effective representation of uncertainty in this case might not be optimal if the model is not tuned to reflect people's error preferences. I propose a method for systematically estimating people's error

preferences using a simple survey and a construct I call acceptability of error. This survey instrument connects classifier evaluation metrics with acceptability and intent to use through the Technology Acceptance Model, and conceptually helps assign costs in classification in accordance with people's preferences for different types of errors. Through face validation of the instrument I found that it is sensitive to differences in people's preferences induced by different types of user interfaces applied to the same problem.

Effective design for user-facing uncertainty is only going to become more important. Someone today might abandon their weight scale, or bus tracking app, or weather forecaster due to too much perceived error. In the future, people will be abandoning smartphone and smartwatch apps for tracking blood sugar, or heartrate variability, or blood pressure. As we push more of these sorts of low-cost sensing and prediction applications into the world, we are introducing more estimates with more error into people's lives (your smartwatch blood pressure will not be as good as a blood pressure cuff). My work aims to help us design systems that bring users' understanding along into this new world.

Table of contents

Acknowledgements

My thanks are not limited to those directly involved in the research I happened to squish into this document. My PhD was fueled by myriad collaborations and side projects (*side project* being slightly retroactive); without the variety of work I was able to conduct at the University of Washington—and the variety of people I was able to conduct it with—I would long since have burnt out. Thus, I first thank the institution of *dub*, and all of the people that make Human–Computer Interaction thrive across the University of Washington. By necessity I have condensed only a subset of my work into this document, but I will try to thank everyone I have collaborated with.

I should thank the two people who helped me get to the point of even considering a PhD. Michael Terry advised me during my Master's at the University of Waterloo, which in a strict technical sense led me to pursue a PhD at the University of Washington. More correctly, however, Mike reminded me that Computer Science is not an inhuman, artless discipline when I was near ready to abandon it. During that time, I was also fortunate to work with Ed Lank, who was then and has been since (over beers at conferences) an endless source of perceptive—if sarcastic—advice about the ins-and-outs of academia.

My PhD advisors, Julie Kientz and Shwetak Patel, have both been tireless advocates for me. On numerous occasions I have discovered a good word has been put in on my behalf by one or the other of them that has opened an important door. They have also put up with a lot: the fact that I have so many people to thank here is in part due to the freedom I was afforded to work on what interested me. At the same time, the occasional poke to consider the bigger picture—especially from Julie—was invaluable.

Several other professors at UW provided wonderful collaborations or conversations during my time there. James Fogarty and I shared many laughs amidst serious and insightful conversations about research, though sadly we never did come up with a project to work on together directly. Jacob Wobbrock was a great source of statistical knowledge and discussion, and was wonderfully efficient to collaborate with. Jeff Heer quickly took on a slightly strange side project that might have seemed difficult to publish, and provided exactly the advice needed to publish it. Jessica Hullman has a knack for asking exactly the right question, and finds and consumes background work at a rate I wish I could emulate.

Sean Munson and Cynthia Matuszek are both great friends turned great collaborators. I dearly miss being in the same city as both of them, as wonderful conversation and myriad research ideas flow easily when either is around. With Sean,

our sensibilities for what makes good research are often in tune, but when they are not I can always anticipate a well-reasoned, passionate argument from him over delicious cocktails. Cynthia is the most intelligent, discerning, and (despite what she might have you believe) *kind* person I have the privilege to call friend. Her insight into virtually any topic you might name is always valuable.

My friends especially kept me sane throughout the slog: Nicholas FitzGerald, Mark Yatskar, and Daniel Perelmen. Games and drinks were always a welcome distraction, even if the puns were terrible. My sister, Rachael, besides being a good friend, always had valuable advice about how to manage time and was always willing to swap stories about the silliness of academia. My brother, Nathaniel, always had a so-bad-it's-good movie recommendation to help distract. My parents continue to and have always supported me in whatever I do, and I love them dearly.

My partner, Barry Nelipowitz, has made the last two and a half years of my PhD (and my time in Seattle) infinitely easier and more enjoyable. His patience with my strange work hours and habits. His ability to make me laugh when I am grumpy. His diligence in poking me to finish that last bit of work, or send one more job application. His wisdom in making me take a break, socialize, talk to other humans. Without him I would be so much less happy and so much less productive.

Finally, I thank the many other wonderful friends and collaborators I have had over the years of my PhD: Gilbert Bernstein, Eun Kyoung Choe, Gregory Nelson, Eric Hekler, Tara Kola, Dan Morris, mc schraefel, Kyle Rector, Jared Bauer, Alex Mariakakis, Hanchuan Li, Nick Jones, Sunny Consolvo, Ben Greenstein, Nathaniel Watson, Jesse Shepherd, Michael Grandner, Patrick Gage Kelley, Saeed Abdullah, Elizabeth Murnane, Kenton O'Hara, James Scott, Mike Massimi, Steve Haroz, Shion Guha, and Pierre Dragicevic. I also thank the various agencies that have funded my work: the Robert Wood Johnson Foundation, the National Sciences and Engineering Research Council of Canada (NSERC), the National Science Foundation (NSF), Google, and the Intel Science and Technology Center for Pervasive Computing (ISTC-PC).

1. Introduction

1.1 BACKGROUND

We are increasingly exposed to sensing and prediction in our daily lives (*How many steps did I take today? How long until my bus shows up? How much do I weigh?*). Uncertainty is both inherent to these systems and usually poorly communicated, when it is communicated at all. Existing interfaces typically give users point estimates of predictions or measurements, omitting more complete representations of uncertainty that could help people make more effective decisions. For example, a person deciding when they should get to the bus stop to wait for their bus probably does not want to know when the bus is most likely to arrive; rather, they might want to know how *early* they should get to the bus stop given their risk tolerance for missing the bus (a time that is virtually always earlier than the most likely arrival time, and which requires some understanding of the error in the prediction to estimate). Instead, existing predictive systems for bus arrival times (e.g., OneBusAway [104]) simply give users a point estimate and hope that they learn to deal with errors over time. But how many times does a person need to miss their bus in order to stop trusting predicted bus arrival times altogether?

Communication of uncertainty—or lack of it—occurs across many different types of sensing and predictive systems. Lim and Dey [61] used surveys to investigate the impact of communicating uncertainty on perceived appropriateness of application behavior in hypothetical context-aware applications. They found that communicating uncertainty can improve impressions of an application so long as the application's certainty is high enough. When uncertainty is too high, communicating uncertainty can make users realize just how bad a system really is. On the other hand, they also found that when a system communicates its uncertainty and has made a mistake, people's impressions of the system may be improved because they might be able to explain how the system could have made a mistake [61]. This suggests that communicating error in sensing applications may help people trust those systems more, or even help prevent abandonment.

However, it is not clear how these results manifest in real-world application use. Are there problems of trust or abandonment in existing sensing and predictive systems that are caused by a lack of communication of uncertainty? If there are, how would we change those systems to communicate uncertainty effectively? With respect to the latter question, work in statistical literacy [30] and medical risk communication [29] has suggested that the use of *frequency-based* (or *discrete*) presentations of uncertainty—e.g., *10/100* instead of *100%*—can improve people's probabilistic reasoning.[1] However, these approaches have only been applied to the communication

1. Where *improve* usually means *make more normative* according to Bayesian axioms of probability [30].

of binary or categorical outcomes (e.g., the probability that a patient actually has a condition given a positive test result [37]—a canonical example used in teaching and assessing Bayesian reasoning). Yet many of the quantities we are interested in in our daily lives are continuous (*How long until my bus arrives? How much do I weigh?*). Is there a way to apply the insights from medical risk communication to communicating continuous outcomes?

Finally, underlying all sensing and predictive systems is a model, even if it is a simple and implicit one, such as the output from a load sensor in a weight scale. At the frontiers of ubiquitous computing research, we are creating more complex models atop simple sensors in order to squeeze more value out of commodity hardware; examples include using low-cost accelerometers to track step count or sleep quality (Fitbit), using microphones for cough tracking [57] or fall detection [72], and using electrical noise and water pressure monitoring to track appliances' water and electricity use [34]. As Lim and Dey [61] argue, before we think about communicating uncertainty in these domains, the sensors should anyway be "good enough". Yet there is no definition of good enough: researchers in ubiquitous computing tend to use a standard of around 85%, but this is just a heuristic. In addition, depending on the application, people may be more concerned about different types of errors (e.g., false positives versus false negatives). For example, it is well-known in information retrieval that people are often more concerned with precision than recall [79].[2] Yet in evaluating novel classifiers in ubiquitous computing, the F1 score remains a common benchmark, which weights precision and recall equally. Are people sensitive to different types of errors in everyday sensing and prediction, such that we should reconsider how we evaluate those systems? And just what is *good enough*, and can we estimate it cheaply?

2. If many documents potentially match a query and your user will not scroll past the first ten, you care more that the few documents they *do* see match their query (precision) than that you return every document that matches (recall) [79].

Problems of trust and abandonment in sensing and predictive systems precipitated by poor communication of uncertainty are only going to get worse. Someone today might abandon their weight scale, or bus tracking app, or weather forecaster due to too much perceived error. In the future, people will be abandoning smartphone and smartwatch apps for tracking blood sugar, or heartrate variability, or blood pressure. As we push more of these sorts of low-cost sensing and prediction applications into the world, we are introducing more estimates with more error into people's lives (your smartwatch blood pressure app will not be as good as a blood pressure cuff), but we are not taking care to bring our users' understanding along into this new world. In this thesis, I ask: do we need to improve communication of uncertainty (**RQ1**, Chapter 3), how should we communicate uncertainty when we do (**RQ2**, Chapter 4), and how could we tune models to match users' preferences with respect to uncertainty (**RQ3**, Chapter 5)?

1.2 RESEARCH QUESTIONS AND CONTRIBUTIONS

1.2.1 RQ1: How does the lack of communication of uncertainty affect trust in a simple, real-world sensing system (body weight scales)?

Ch 3 describes my work on the effect of communicating uncertainty on trust in home weight scales.

Building on the work of Lim and Dey [61], I wanted to know how uncertainty affected the use of a sensing application in the real world. I wanted a sensing system that was simple, ubiquitous, and which communicated a concrete quantity people could reasonably be expected to understand. If a lack of communication of uncertainty in such a context creates problems of trust or abandonment, we should expect more complex systems to be even more problematic. Thus, I started with the simplest, most ubiquitous sensing system I could think of: the body weight scale. This sensing system has been with us for at least 100 years, and yet it maintains at the moment of measurement about the simplest feedback interface possible: it tells you your estimated weight (and no consumer scales I am aware of communicate the error of that measurement). At the same time, it is a very useful sensor—several studies have shown that frequent weigh-ins help maintain weight loss [96,102]—and ostensibly an easy-to-understand one.

I wanted to know, qualitatively, if and how the lack of communication of uncertainty affected trust in actual users. I coded online reviews of scales for discussions of uncertainty and error, and found that negative reviewers often (about 26% of the time) perceived their scale to have too much error. However, many of the problems with scale measurements people described reflected their lack of statistical sophistication more than a problem with the scale. For example, a biased body weight scale is still quite useful for tracking relative changes in weight over time, so long as its variance is low[3]—yet people will try to judge how well-calibrated their scale is against another (say their doctor's), a measure of bias, to judge how suitable a scale is for tracking their weight. These findings were corroborated by interviews with experts (practitioners who work with people trying to change their weight) and by an online survey of existing scale users, in which I found that people who have a better understanding of weight variability trust their scales more.

3. If my scale consistently overestimates my weight by 2lbs (a bias), I can still use it to track relative change: differences in weights measured on the scale will be accurate even if the particular numbers are not.

Contributions. My results extend the existing literature on intelligibility that used hypothetical surveys [61] by demonstrating evidence for decreased trust (and possible abandonment) of a simple, real-world sensing system due to misunderstandings of uncertainty. From my qualitative investigations, I also contribute design principles and a new weight scale design for communicating uncertainty more effectively, including principles like *avoiding false precision* and *accounting for and explaining known biases through modeling*.

1.2.2 RQ2: Can we build visualizations of uncertainty in continuous predictions (bus arrival times) that people do understand?

Ch 4 describes my work on communicating probabilistic predictions in realtime transit prediction.

It is not enough to note that uncertainty is not communicated effectively in existing systems; I wanted to understand how to communicate uncertainty in a way that people grasp more viscerally. While discrete representations of uncertainty have been successful in medical risk communication [29]—where outcomes are typically binary—could these approaches be applied successfully for continuous predictions or estimates?

To answer this question, I turned to another domain—realtime transit prediction, and specifically users of the OneBusAway realtime transit prediction application in Seattle [104]. This domain presents additional challenges in communicating uncertainty: OneBusAway is a mobile app, meaning that people are using it to make quick decisions (decisions are time-constrained) and are typically looking at multiple predictions simultaneously on a small smartphone screen (the display is space-constrained). The best recommendations from the literature for communication probabilistic predictions involve dual plots of a probability density and the cumulative distribution function of an estimate [39], which is not possible in such a space-constrained environment. Thus, I was interested in developing a visualization that fit into this time- and space-constrained context and which was easy for people to understand (perhaps by capitalizing on frequency-based representations).

I first conducted a survey of existing users of OneBusAway in order to determine if they had unmet needs that could be addressed through probabilistic prediction. I found that they did: the most desired features included a need for *status probability* (how likely is it my bus has gone by?) and a sense of *prediction variance*. Through iterative design and paper prototyping with users, I developed a new OneBusAway interface to address these needs. I also developed a novel visualization of a probability distribution, the *quantile dotplot*, that uses a frequency-based representation to make interval estimation easy. In an online survey of existing OneBusAway users, I found that quantile dotplots decreased the variance of people's probabilistic estimates by about 15% compared to a probability density plot.

Contributions. My results document specific needs of users of realtime transit prediction applications that could be met through probabilistic prediction. I contribute a redesigned transit prediction interface to address those needs, as well as a novel discrete visualization, the *quantile dotplot*, that extends successes in discrete communication of uncertainty in binary outcomes [29] to continuous outcomes. If my work in weight scales asks what happens when we get it wrong, this work aims to build an understanding of how to get it right.

1.2.3 RQ3: Can we tune models to people's error preferences in a simple, lightweight way?

It is not enough to slap an effective visualization on top of whatever model exists in a predictive system. In even conceptually simple prediction tasks—*is it going to rain today?*—people exhibit preferences for different types of errors. In precipitation prediction, this manifests as *wet bias* [85]—the tendency of weather forecasters to over-predict the probability of rain to reduce the chance that people do not prepare for rain and then blame the forecaster when they are rained on. Even an effective representation of uncertainty in this case might not be optimal if the model is not tuned to reflect people's error preferences. Given known costs for each type of error, *cost-sensitive classification*[4] can be employed to fit a model that makes predictions that reflect error preferences.

Ch 5 describes my work on estimating acceptability of error in several sensing and prediction domains.

4. For an overview of cost-sensitive classification, see [23]; for an example of a method to employ cost-sensitive classification in arbitary classifiers, see MetaCost [21].

In a business context, costs of different types of errors may be determined financially (i.e., the costs are in predicted gains or losses to the business due to correct or incorrect predictions). However, in a user-facing context that may not directly involve financial stakes, how do we assign costs? Arguably, part of the cost of errors of a user-facing system lies in whether the user finds that system suitably *acceptable* for its task, and consequently whether they consider it useful enough to use. Variants of the Technology Acceptance Model (TAM) have commonly been used to estimate constructs like *usefulness* in order to predict *intent to use* [98,99]. Thus, I propose using a modified variant of the TAM2 [98], including a new construct I call *acceptability of error*, to estimate people's error preferences for classifiers employed in user-facing sensing and prediction. Specifically, I propose a simple survey instrument that elicits acceptability of error over a range of possible values of precision and recall for an application, and then estimates a weighted mean of precision and recall that represents users' error preferences in that application.[5]

5. This allows measures like weighted F-score, already familiar to UbiComp researchers, to be used to evaluate classifiers. I view this as a user-centric approach to cost-sensitive classification (where users are applied researchers in UbiComp), though a full cost matrix could also be estimated and fed into a cost-sensitive classifier.

I conducted a face validation of my survey instrument in an online survey that used four different applications inspired by the UbiComp literature. I found that the instrument was sensitive to expected differences in users' error preferences, including differences between two user interfaces for the same underlying classifier. In another survey, I demonstrate the predictive validity of the instrument by predicting the acceptability of error of weather forecasters people use based on their real-world error rates.

Contributions. My results help formalize the notion of acceptability of error, and demonstrate how it can be associated to traditional measures of classifier error. I also provide initial validation for a survey instrument that developers of inference-based systems can use to help identify acceptable levels of classifier performance before expending the effort to build the systems.

1.3 THESIS STATEMENT

Through my research questions I aim to tackle the full *stack* involved in communicating uncertainty: from user understanding at the top, through representations people can understand, through models that reflect their preferences. The remainder of this thesis is organized around the above research questions in support of the following thesis claims:

T1. By using richer probabilistic models of the uncertainty in predictive estimates—conveyed to users in ways that are relevant to their goals, that are in representations they can understand (e.g., by using frequencies and concrete examples instead of probabilities), and that are sensitive to the types of errors they care most about (e.g. false positives versus false negatives)—we can **A)** improve users' acceptability of error in sensing and predictive systems, **B)** improve users' trust in those systems, and **C)** make it easier for users to answer questions they care about when using those systems.

Ch 3 presents work on understanding of uncertainty in weight and some evidence for **T1B**. Ch 4 builds on this work in another domain, with evidence for **T1B** and **T1C**. Ch 5 suggests a way to estimate acceptability of error, towards **T1A**.

T2. Using simple survey-based methods that elicit acceptability of error across a space of hypothetical errors in an application, we can A) estimate how much users care about different types of errors in a predictive application (even without ground truth), and B) express those estimates in the domain language of practitioners building predictive systems such that they can be easily adopted (e.g. as a weighted mean of precision and recall).

Ch 5 presents a model for estimating (**T2A**) acceptability of error expressed in terms of precision and recall (**T2B**).

1.4 THESIS ORGANIZATION

Research question	Thesis claims	Addressed in…
RQ1: How does the lack of communication of uncertainty affect trust in a simple, real-world sensing system (body weight scales)?	**T1B**, **T1C**	**CH 3**, through a qualitative study of online reviews of scales written by scale users and through a survey of scale users.
RQ2: Can we build visualizations of uncertainty in continuous predictions (bus arrival times) that people do understand?	**T1C**	**CH 4**, through redesign.of a bus prediction interface and a study showing people's improved estimates of probability using quantile dotplots.
RQ3: Can we tune models to people's error preferences in a simple, lightweight way?	**T1A**, **T2A**, **T2B**	**CH 5**, through the development of an acceptability of error survey instrument and surveys testing its face and predictive validity.

2. Related work

2.1 END-USER PERCEPTIONS OF UNCERTAINTY AND ACCEPTABILITY OF ERROR

2.1.1 Intelligibility

A growing body of work in Human–Computer Interaction (HCI) and Ubiquitous Computing (Ubicomp) has involved investigations of the *intelligibility* of user interfaces: how transparent the reasoning or certainty of these systems are to users [60,61]. The effects of intelligibility seem to be application-dependent: displaying uncertainty sometimes has positive [1] or negative [81] effects on task performance. In a study of several hypothetical context-aware systems, Lim and Dey [61] found that making the certainty of a system visible to users—for example, as a confidence region in location-aware systems—can improve users' perceptions of the accuracy and appropriateness of a system, so long as the accuracy is good enough. Another way of looking at this is that, in a system with low intelligibility (e.g., where accuracy is not conveyed effectively to the user), a user has less information on which to judge accuracy, so their judgments naturally tend to be more uniform across varying levels of accuracy. However, if that same system is more intelligible—more clearly conveys its accuracy—the user's opinion of that system goes down if the system is actually less accurate than they thought and goes up if it is more accurate than they thought [61].

However, in the context of an inference-based system, it is not clear what components of accuracy contribute to these judgments. For example, it is well-established in information retrieval literature that the unweighted F1 score is inadequate for many applications, since users may be more concerned (for example) with precision than recall [59, 79]. Yet, we still commonly use F1 score in evaluating classifiers in many user-facing applications. It is worth asking whether or not we can more systematically estimate the individual effects of precision and recall on acceptability of error.

2.1.2 Preference and choice modeling

Behavioral economics and the marketing have developed models for estimating how different properties of some product contribute to ratings of that product or to the probability that a person would buy that product (or to its utility), areas referred to as *stated preference modelling* and *choice modeling* [10,65]. Merino-Castello provides an overview of several different type of preference modeling [65], including *multi-attribute valuation*, in which multivariate regressions are conducted to estimate the effects of different properties of a product on either stated preferences for that product or on choices between alternatives. In the case of preferences, this

is called *conjoint analysis*, and specifically in the case of predicting stated ratings of goods and services, it is called *contingent rating* [65] (presumably because people's ratings of goods and services are assumed to be contingent upon the properties of those goods and services). While the approach discussed therein relies on Ordinary Least Squares (OLS) regression, I take a similar approach using (ordinal) logistic regression (invoking fewer assumptions about the nature of the preference scale) in Chapter 5 to estimate acceptability of error as a function of precision and recall.

Preference and choice modelling can also be distinguished as being based on *stated preferences* (as above) or *revealed preferences* [97]. In the former case, responses to hypothetical scenarios are used to prompt for preference ratings (as in contingent rating) or by directly asking for a preferred alternative amongst a set (*discrete choice modeling*, which then uses properties of the choices to predict the probability of an alternative being chosen [6]—to answer questions like, is a red or blue car more likely to be chosen?). Revealed preferences, by contrast, are derived from real-world behavior (such as cars actually purchased). While this type of data better reflects likely behavior than stated preferences, it is more difficult to use it to predict preferences amongst non-existent (e.g. hypothetical) goods and services, or to systematically evaluate the effects of different properties of those goods and services (because in real-world data those properties may be correlated) [65]. A similar problem presents itself when trying to analyze the hypothetical accuracy of user-facing systems: often we wish to know how good our system needs to be (not how good it is now), and as stated above, we may wish to know which types of errors are particularly relevant to users—but these errors may be correlated (e.g. precision and recall of a real-world system may be similar). There is therefore an advantage to developing a technique that can estimate real-world use from stated preferences, perhaps by combining stated preferences with some ground truth data; Section 5.5 proposes such a model for systematically estimating the individual effects of precision and recall on the acceptability of error in inference-based applications in the domain of weather forecasting.

2.1.3 Evaluating acceptability of error in HCI and Ubicomp

In addition, given a highly intelligible system with acceptable levels of accuracy, it still behooves us to ask whether users find it to be useful. To that end, I use a variant of the Technology Acceptance Model (TAM) to validate my measure of acceptability of error. TAM is a well-studied method for predicting technology acceptance, originally proposed for use in the workplace [19]. Since then, numerous variants of TAM have been proposed [98,99], and it has been applied to contexts outside the workplace, such as e-commerce [70] and consumer health technology [67]. The core constructs of TAM include perceived ease of use, perceived usefulness, and intent

to use a technology, which have been shown to predict real-world use [19,70,98]. In this work, I adopt a variant of the TAM2 [98], which includes a construct called output quality—how well a system performs the tasks it is designed for—which I believe to be related to acceptability of error in ubicomp systems.

The development of methods to evaluate ubicomp systems that use sensing and inference has been a popular topic within the last decade. Several frameworks have been proposed for evaluating sensing and ubicomp systems [5,53,84]. These frameworks aim for a holistic evaluation, whereas I explicitly look toward a method for assessing the acceptability of error. Others call for evaluating ubicomp technologies through in-situ deployment studies of built systems [80]. This can be a very useful method to assess the acceptability of error, and studies of applications that use sensing have been able to evaluate the acceptability of error of an already built system within the context of use (e.g., [15]). These deployments are very resource-intensive, however, and thus I aim to reduce the overhead of assessing the acceptability of error before such systems are built. Finally, other researchers have proposed methods of formative assessment of ubicomp systems through the concepts of sensor proxies [11] and experience sampling [14], but these methods still require in person interaction with participants, and do not provide explicit guidance on the acceptability of error of inference systems. I believe my method can complement these existing approaches. In particular, by modeling acceptability of error as a function of measures familiar to developers of machine learning applications—and by expressing its results as an objective function that can be optimized by learning processes—I provide a model of acceptability of error that is expressed in the domain language of the experts who build these systems.

2.2 REPRESENTING UNCERTAINTY AND END-USER UNDERSTANDING OF UNCERTAINTY

2.2.1 Risk management

Leiss [58] notes a progression in the field of communicating risk management that went from what he calls *Phase I* approaches to risk management (quantifying objective risk to the public, which presented problems of understanding—experts' models and measurements of risk may not be well-communicated to the public) to *Phase II* approaches (essentially, attempting to address the gap of understanding by better communicating objective risk measures and convincing the public of their correctness), to *Phase III* (making people partners is assessing risk). The key insights in this space were first that the public did not accept attempts to be *persuaded* to accept experts' conceptions of risk; instead, it was necessary to incorporate public perspectives into risk management [58]. The lesson for representing uncertainty in user-facing systems is that it's not enough simply to quantify uncertainty for

users, but also to incorporate some sense of users' understanding and desires to effectively communicate risk and uncertainty—the thrust of **T1**. At the same time, this suggests a need for the models underlying predictive systems to better reflect the trade-offs of concern to users—the thrust of **T2**.

2.2.2 Trust and decision-making when uncertainty is communicated

Lab studies have found that communicating uncertainty (or not) can affect the trust that people have in computing systems, and also can affect the types of decisions people make when uncertainty is present. Jung *et al.* [44] found that displaying the estimated remaining range of an electric vehicle as a gradient plot (i.e., with uncertainty) reduced range anxiety in a driving task compared to a single point estimate. Joslyn & LeClerc [42] found that displaying uncertainty in weather predictions can lead to more optimal decision-making and trust in a forecast. When asked to make decisions about whether to salt roads (given a virtual budget, cost for salting, and cost for failing to salt when they should have), people made more optimal decisions when given point estimates with probabilistic information. Subjects with access to probabilistic information even made more optimal decisions than subjects who were explicitly told the optimal decision based on a cot-benefit analysis. While the decision suggested by a cost-benefit analysis will give the best choice on average, always applying the decision will sometimes lead people to take precautions that seem unnecessary (e.g., salting the roads when the weather ultimately does not require it). After experiencing a such few errors, people may begin to distrust the strategy and ignore the suggested course of action [42]. Probabilistic information, on the other hand, provides a more transparent form of information for decision making, leading to greater trust.

2.2.3 Representing uncertainty

Though it is generally accepted that individuals draw more informed conclusions when they integrate a notion of variability into their judgments [90], most individuals have trouble conceiving of and integrating probability into their decisions [94]. Taken with the lessons from risk management, we have an opportunity to better communicate uncertainty not only so that individuals understand it, but so that it aligns with their goals.

Additional related work on visualizing uncertainty is covered in-context in Section 4.2 while establishing design guidelines for visualizing uncertainty.

There has been extensive work in visualizing uncertainty [40,41,62,83,108] (see [108] for an overview of several techniques in scientific visualization). Much of this work focuses on scientific visualization and perceptual trade-offs of different encodings for representing probabilities (e.g. the use of shading versus other encodings to visualize probability distributions of point estimates or continuously varying uncertainty [40,83]). While this work offers low-level guidance in representation,

it does not address issues of low graphical literacy in the broader population [28], which presents a problem when designing visualizations of uncertainty for non-scientific audiences. The most relevant area of work seems to be the communication of health treatment risk to patients; studies in that area have found (for example) that visualization of risk (as discrete pictograms representing a population) can reduce cognitive biases in estimating risk [29,41]. In Chapter 4 I explored many traditional graphical approaches to representing probability during design iteration, such as interval based methods (e.g., textual or visualized prediction intervals), as well as more complete visual representations of probability distributions, such as probability density functions (PDFs) or cumulative distribution functions (CDFs); however, it is not clear that these approaches will be most effective for lay audiences. For example, there is evidence that interval-based methods like error bars are difficult even for experts to interpret [4], and non-experts do not likely encounter representations like PDFs and CDFs regularly, suggesting a need to explore this design space more broadly.

One alternative route is visualizations of discrete scenarios or events: previous research in decision-making science indicates that people's ability to reason about uncertain information is improved when that information is framed as natural frequencies (relative counts) rather than probabilities [30]. Many common visual uncertainty representations have focused on depicting probabilities, while non-visual studies of the power of natural frequencies have largely been restricted to controlled lab studies. In static representations, these approaches have also tended to involve binary measures (as in medical diagnostic tests [29]) rather than continuous measures. One avenue I explore in Chapter 4 is to convey uncertainty in a static, discrete visualization for a continuous measure (introducing a technique I call *quantile dotplots*). The frequency approach also motivates the formulation of hypothetical scenarios as counts of discrete events when developing the acceptability of error instrument described in Chapter 5.

2.2.4 Natural language representation of uncertainty

Other work has looked at using natural-language generation to describe inferences in health data [73,91] as a way to improve human inference. I believe this approach may be promising for other user-facing application areas, particularly as natural language generation can improve *expert* inference over strictly graphical methods [4], and layperson graphical literacy is low [28]. A systematic understanding of lay statistical vocabulary is essential to such an approach. Researchers have tried to quantify *words of estimative probability* by having people assign numerical probabilities to words like 'likely', 'uncertain', 'impossible', and so on [51]. Similarly,

confusion around measurement descriptions such as ‘precision’ and ‘accuracy’ has been explored in science education [92] and in specific scientific domains [101], but I am not aware of similar investigations of lay understanding of such words, despite their frequent use in product descriptions and consumer reviews.

3. Test case for understanding of uncertainty: body weight

3.1 INTRODUCTION

I wanted to begin my investigation into problems with communicating uncertainty in everyday sensing with a simple and ubiquitous sensor, one that is ostensibly straightforward for people to understand. I chose to examine the bathroom scale, which is arguably the most ubiquitous health sensor of all—and which is certainly the most ubiquitous tool for diagnosing and managing weight issues. Several studies have shown that frequent weigh-ins help maintain weight loss [96,102]; despite this, people who are watching their weight often have a marked aversion to stepping on the scale [22]. I believe that some of this resistance comes from the design of the scale's interface, and particularly its failure to communicate uncertainty.

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 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. Due to both its ubiquity and the (seeming) simplicity and familiarity of the data it presents, as well as the simplicity of the way it typically presents data (as single measurements without context or uncertainty), the weight scale offers an excellent test case for the claims of **T1**, by illuminating problems with trust in data caused by unsophisticated presentation of seemingly simple data.

I believe there are several issues with current scales interfaces. For example, digital scale readouts convey an unrealistic level of precision, negatively affecting user perception. In an online survey of over 800 scale users, I found that respondents with less understanding of how weight fluctuates during the day were less likely to trust their scales (**T1B**). This is exacerbated by the fact that the scale interface makes no attempt to inform users about how weight fluctuates, or what level of precision to expect. This 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

This chapter is based on work published at *UbiComp 2013* with Dan Morris, mc schraefel, and Julie Kientz [48].

This chapter centers on **RQ1:** *How does the lack of communication of uncertainty affect trust in a simple, real-world sensing system (body weight scales)?*

For reference, **T1** is also repeated here:

By using richer probabilistic models of the uncertainty in predictive estimates—conveyed to users in ways that are relevant to their goals, that are in representations they can understand (e.g., by using frequencies and concrete examples instead of probabilities), and that are sensitive to the types of errors they care most about (e.g. false positives versus false negatives)—we can **A)** *improve users' acceptability of error in sensing and predictive systems,* **B)** *improve users' trust in those systems, and* **C)** *make it easier for users to answer questions they care about when using those systems.*

as body fat estimators, thermometers, pedometers, and blood pressure cuffs—this work provides a foundation for future design in this broader space.

3.2 BACKGROUND ON WEIGHT MANAGEMENT AND SCALES

As links among obesity, mortality, and other health conditions have become clear [2,35], weight management has become a key part of health practice. Obesity is clinically defined in terms of weight and Body Mass Index (BMI) [9,89]; 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 [96,106]. 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 [52,102]. Actual approaches to reducing weight are most commonly associated with calorie restriction and increased physical activity [54,66]—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 has been directly associated with increased mortality [20,82]. 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 [20,53].

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 [13,76]. 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.

3.3 ONLINE REVIEWS STUDY

In order to first assess the effects of current weight scale presentations on trust and acceptability, I began by investigating users' perceptions of weight scale data as expressed in 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 (i.e. trust and acceptability of accuracy)? and 3) What terminology do consumers use to express these expectations? This last point will become particularly relevant when discussing how to improve the presentation of that data.

I 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 I convert it to a helpfulness score by taking the lower bound of its 95% binomial confidence interval.

From a corpus of 1084 reviews, I selected those with at least one helpfulness rating (855 reviews). Of these, I 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[1] used affinity diagramming to identify recurrent themes within this subset around users' understanding of precision, accuracy, and uncertainty. I 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 I used frequency profiling [77] 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.

1. Two other authors of the original paper [48]. and I.

3.3.1 Results

...1 Trend focus versus 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 inaccurate data, so long as they can establish a baseline from which to observe change: these users understand that given inaccurate but reasonably precise data (i.e. data that has a consistent bias), trends are still meaningful even if the particular numbers are inaccurate. 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.

…2 Vocabulary and terminology

In total, 68 of the 100 reviews I coded discussed issues around accuracy, precision, or uncertainty. To get a sense of the vocabulary used to express these concepts, I counted the number of reviews containing various words and their derivatives (I 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). Notably, 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. I also observed a strong preference for the use of the term 'accuracy' to refer broadly to issues of measurement uncertainty. A more systematic investigation of vocabulary for expressing uncertainty is clearly warranted in order to be able to effectively communicate these concepts to end-users; an initial pass at such an investigation is later in this chapter.

3.4 EXPERT INTERVIEWS

I interviewed four experts on weight change to validate the findings from the 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.

I 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. I used affinity diagramming of transcripts to identify high-level themes, discussed below.

3.4.1 Results

...1 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".

...2 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 the 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.*

...3 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".

...4 Regular weighing still has significant value

Despite the potential issues with weighing the 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 the 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, the 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'".

...5 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 [60]—may have strong impact in the weight space, further reinforcing a need to understand how laypeople use statistical vocabulary.

3.5 WEIGHT TRACKING STUDY

The results of the online reviews study and expert interviews support the 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. I was unable to find studies of within-day weight fluctuation in the literature (weight change is typically studied between days). Furthermore, consultations with physicians and dieticians 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, I devised a study to gather data on within-day weight fluctuation. I 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 my 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—that a lack of communication of any sort of reasonable model of weight data hampers trust and acceptability of accuracy of this data (**T1**).

Figure 3.1. Screenshot of the 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.

I used a journaling approach to collect multiple weigh-ins from users on a mobile web app (Figure 3.1). I recruited within Microsoft Research Redmond (MSR) via a departmental email list and on weight-related Internet forums. For participants within MSR, I placed 10 digital weight scales of the same model throughout the 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 3.1). I asked participants to weigh themselves at least 3 times daily for a period of at least 10 days, spanning two weekends, and to use the web app to report their weight immediately after weighing. In addition to the user's current weight, the 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, I had data from 23 participants (69% male): 17 internal to the organization and 6 external. Participants weighed themselves an average of 28.8 times (sd=23.8, min=6, max=109); 15 participants provided at least 20 measurements. Mean weight among participants was 168.2 lbs (sd=8.5), mean age was 32.5 (sd=9.4).

3.5.1 Results

...1 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. I used a mixed-effects 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 the 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 3.1:

Component	Effect (lbs)	SE		
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 3.1 Effects of weighing conditions on weight.

On average, being partially clothed increased weight by 0.85 lbs and being fully clothed increased weight by 2.17 lbs. The model also estimates an offset for each scale from the mean 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) [32], supporting the model's validity.

...2 Within-day weight variation

To estimate typical within-day weight variation, I considered all instances of any participant submitting at least 3 weigh-ins in a calendar day. I then calculated the difference between the maximum and minimum recorded weight for each day; I call this the within-day range. The model of clothing and scale effects also allows me to derive an adjusted weight for each weight. I 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, I 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, I report both (Table 3.2, Figure 3.2a). The mean within-day range was 3.60 lbs (2.72 lbs adjusted), validating the experts' estimates of about 3–5 lbs.

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 3.2 Unadjusted and adjusted within-day weight ranges.

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 the 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 and loss of trust amongst people who weigh

themselves often (or with different scales) but who do not fully understand these sources of weight change.

...3 Weight range by mean weight

I 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. However, I 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 I saw no evidence for such a relationship, it is worth noting that I 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 the sample.

...4 Focus & limitations

I stress that the regression used in the study was only to approximate the fluctuation in weight measurements, as the primary focus of this work is on examining the appropriateness of instantaneous measurements of weight (without uncertainty) from an end-user perspective. That is, the physiological influences on weight fluctuation (menstrual cycle, salt intake, etc.) are not in this scope: I wanted to know what people's weight fluctuations look like to them, regardless of what caused them. The focus on fluctuation—not on causes of fluctuation—precisely complements the observation that scales do not use or present any of this potentially explanatory information either, to the determine of users' trust and understanding.

3.6 SCALE PERCEPTIONS SURVEY

A pervasive theme throughout this investigation was users' struggle to understand and account for fluctuations in data: both in product reviews and expert interviews, I encountered mismatches between the magnitude of reactions to weight change and the actual significance of that change, given the knowledge of actual weight fluctuation derived from the weight tracking study. I 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? This approach acts as an initial investigation into the potential gains in trust and acceptability that may be possible by improving the scale user interface (**T1**). Noting the inconsistent use of statistical vocabulary by product reviewers, I was also interested in establishing a common lay vocabulary for scale properties like accuracy and reliability; this vocabulary might be useful in certain educational or natural language presentations of this data.

I recruited via mailing lists within MSR, 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. I also invited participants from the weight tracking study to complete an exit survey that included questions 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.

3.6.1 Results

Of 892 total respondents, 18 were participants in the weight tracking study and 30 were internal to MSR. Of the 861 others, 716 were recruited via E4, who advertised the 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.

...1 Understanding of within-day weight fluctuation

To estimate respondents' understanding of typical daily weight fluctuation, the survey 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.*

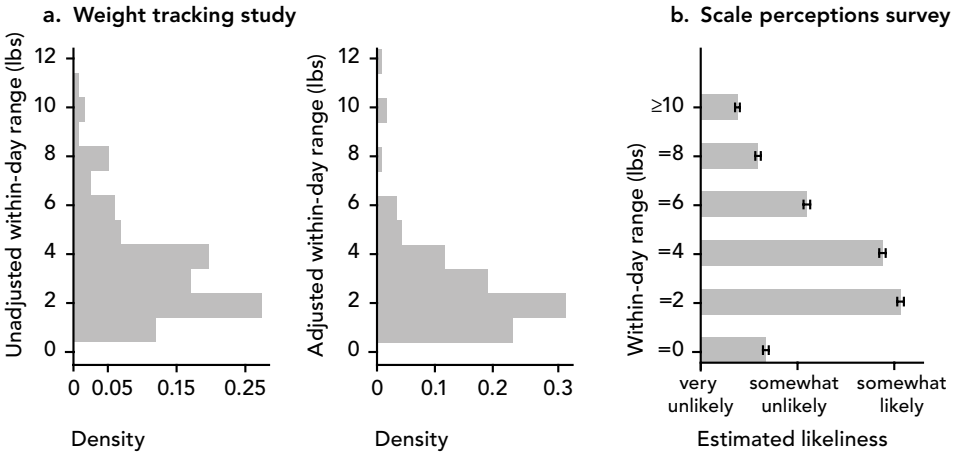


Figure 3.2. a) From the weight tracking study: histograms of within-day weight ranges (max – min weight within a day) before and after adjustment for weigh-in conditions.

b) For comparison, from the scale perceptions survey: respondents' estimated likeliness of various within-day weight ranges.

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, I wanted respondents to indicate their expected distribution of within-day weight ranges. Results of these questions are shown in Figure 3.2b alongside the distribution of within-day ranges from the weight tracking study.

Respondents' estimations of within-day weight range were generally good: the shape of their average estimated distribution is similar to the observed distribution. Respondents tended to place 2 lbs or 4 lbs as the most likely weight range, close to the 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.

...2 Weight fluctuation knowledge and weight data perception

To compare responses between respondents who had a more or less accurate understanding of daily weight fluctuation, I categorized their likeliness estimates into accurate and inaccurate estimates. Accurate estimates were those that: (1)

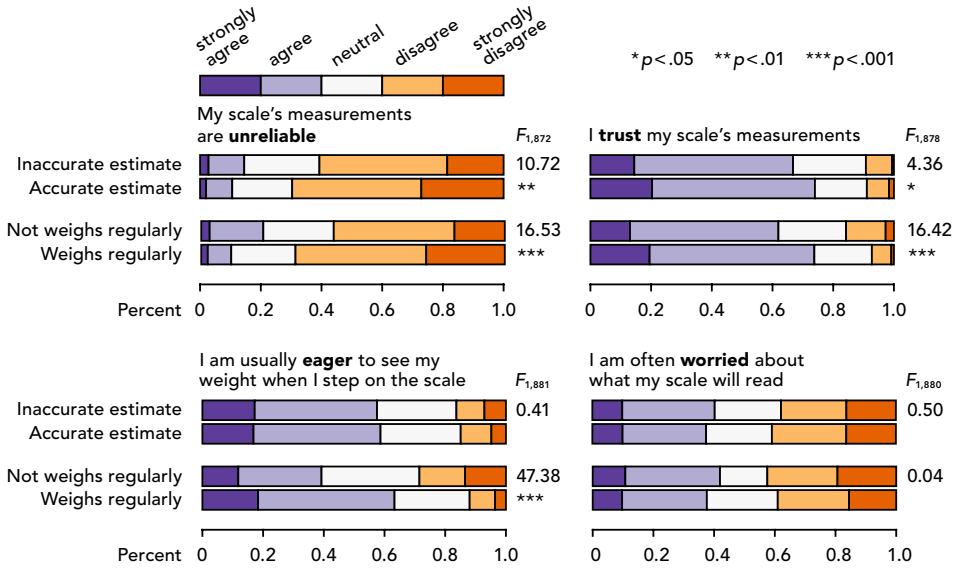


Figure 3.3. Results of the four Likert-scale questions on scale attitudes, broken down by the quality of the respondents' estimation of within-day weight fluctuation and by whether or not respondents weighed themselves regularly.

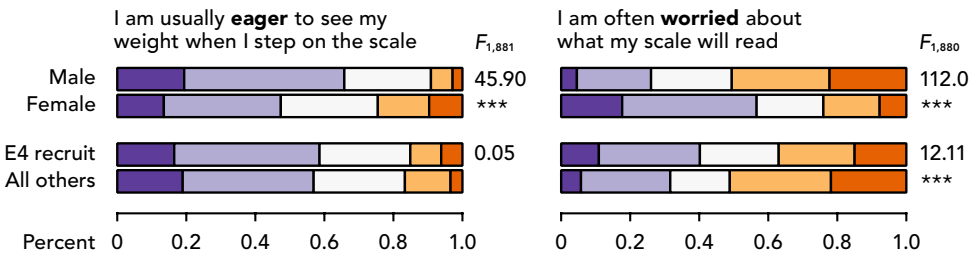


Figure 3.4. Gender and recruitment strategy differences.

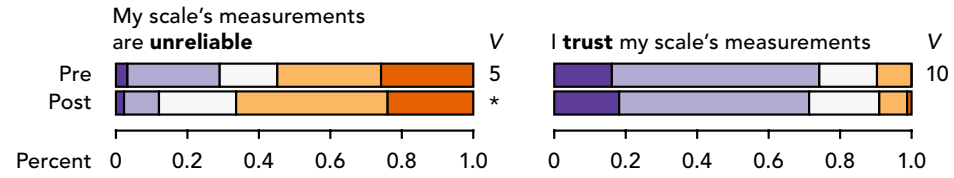


Figure 3.5. Likert questions on scale attitudes asked before and after participation in the weight tracking study (N=15).

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*. I did not factor the 6 lbs range into the 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 I 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, I asked respondents four Likert-item questions on their attitudes toward scales: unreliability, trust, worry, and eagerness (Figure 3.3).

To analyze the Likert data, I used the Aligned-Rank Transform (ART), which allows nonparametric testing of multiple factors with an ANOVA [107]. I 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 as factors in the analysis. I included the latter factor because regular weigh-ins improve weight change outcomes [96,106], and I was curious if it was associated with perceptions of weight data. I 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 3.3).

I 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 [22]. Origin had a significant effect on *worried* (Figure 3.4):

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

those recruited via E4 were more worried at what the scale would read, but had no other significant differences from other respondents.

I also asked the first two Likert questions (*unreliability* and *trust*) to participants in the weight tracking study in a pre-study survey (Figure 3.5). Because these participants also filled out the weight perceptions survey as an exit survey, I was able to see whether tracking their weight and seeing their graph of daily fluctuation had an effect on their attitudes towards scales (note that this is purely a correlative exercise, as the weight tracking study was designed primarily to collect weight data and did not have a control without weight feedback). 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.

...3 Common vocabulary

Following the discovery from the online reviews study that people have varying vocabulary when it comes to expressing uncertainty in scale measurements, I sought to find a common vocabulary to communicate concepts like the accuracy or precision of a scale to consumers. In particular, I was interested in scale accuracy, measurement reliability (both internal and external), and scale readability (also called resolution). I created four definitions to reflect these concepts as applied to scales and refined them through survey piloting (Table 3.3).

I 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 the 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, and repeatability. All of these words (or derivatives) appeared in a subset of the product reviews I examined and were used by reviewers to refer to some property of the data, although not necessarily in a consistent manner. Table 3.3 shows the results of the vocabulary questions, with the most common definition for each word highlighted.

3.6.2 Discussion

The survey 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 [96, 106].

Nearly 40% of the survey sample did not have accurate estimates of within-day weight range. 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 this effect was independent of whether or not those people weighed regularly. While it is also difficult to assign causality here, when considered alongside the other results from this chapter, it is suggestive: online product reviewers are often overly concerned about fluctuations in weight that the weight tracking study suggests are typical, and experts find it essential to educate people about daily weight fluctuations to increase compliance and allay concerns. In this study, I found that people with greater understanding of those fluctuations 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. When considered alongside the lack of any reflection of a more sophisticated model of weight data in current weight scale interfaces, we might ask: could we improve user understanding of weight data through improving the interface, and therefore improve trust and acceptability at the same time?

3.7 DESIGN RECOMMENDATIONS

In this section, I synthesize design recommendations for weight scales based on all of the study results in this chapter. I believe that many of these recommendations could be generalized to other single-point sensors as well.

3.7.1 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 I believe is a more widely understood term. Interestingly, in statistics, 'precision' is often given this definition [8], 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 distinct concepts. The confusion between 'precision' and 'accuracy' has been recognized in other domains [92]; these results suggest that the confusion may be primarily one-way: 'precision' is often used to mean 'accuracy', but the opposite may not be true.

I consider these recommendations an important starting point for exploring natural language feedback techniques in the design of weight scales and similar sensing devices.

3.7.2 Reflect data uncertainty

Traditional scales do not adequately convey uncertainty: as found in online product reviewers' concerns about accuracy—suggesting scales are not conveying accuracy and precision well, or when either is an appropriate concern—and in experts' discussion of emotional reactions to weight fluctuation—"the horrors". Reinforcing these results, the online survey found that greater knowledge of weight fluctuation was associated with higher trust in scale data. Here I offer specific recommendations to improve how scales reflect uncertainty based on these results.

...1 Avoid false precision in single-point measurements

Digital scales, even inexpensive ones, typically have quite fine resolutions (0.2 lbs or less). However, the 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. I 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 experts, in which people identify with a particular weight they want to be rather than focusing on healthy change. I 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.

...2 Adopt an explicit model of weighing frequency

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 the weight tracking study above, or from the user themself) 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 the product reviewers and reported by experts and explicitly enforces data use patterns advocated by the 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.

...3 Educate users about uncertainty

In the online survey, respondents with a greater knowledge of within-day weight fluctuation had greater trust in their scales' accuracy. Similarly, in the weight tracking study, participants' perceptions of the unreliability of their scales decreased after seeing regular graphs of their daily weight fluctuation. Experts also stressed the importance of education to adoption of healthy patterns of scale use. I 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

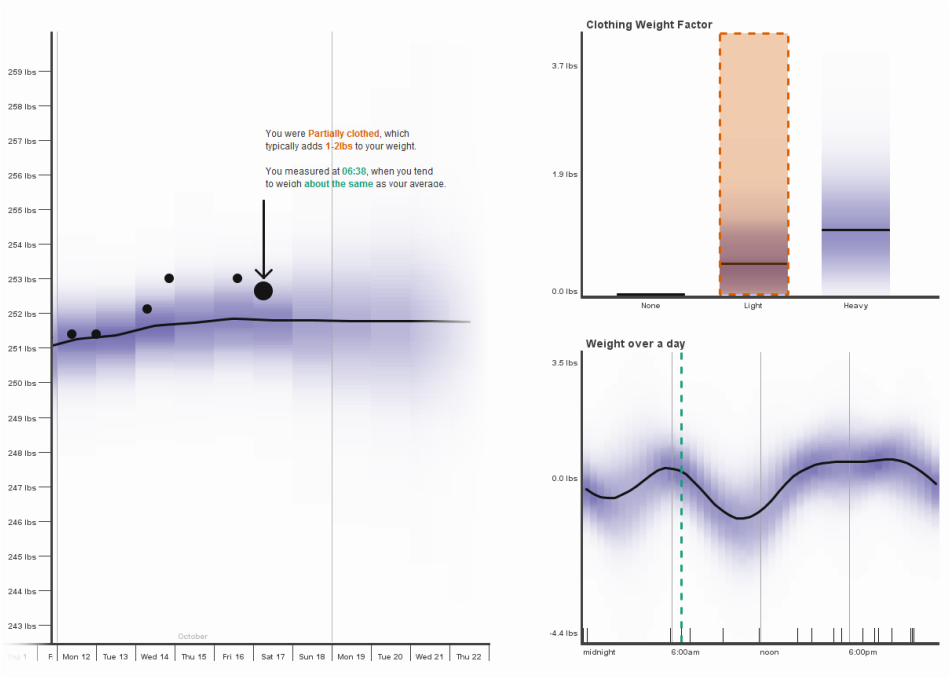


Figure 3.6. Example of redesigned weight scale interface.

within 3 lbs since weight typically fluctuates about that much during the day", or may involve graphical depictions of weight variability. Indeed, the lightweight graphing approach from 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 [108] 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 [40]) may be particularly applicable to the weight domain and to implementing the above recommendations.

3.8 REDESIGNED SCALE

In ongoing work, I have been developing a redesigned weight scale based on the recommendations developed above.[2] I have conducted iterative designed and testing of a modified scale interface that presents uncertainty to its users.

2. Along with Alex Mariakakis, Hanchuan Li, and Nick Jones. This work has not yet been published.

3.8.1 Interface overview

I used a scale with a serial port interface connected to a laptop to replace the default single-data-point display with a display that consists of three parts (Figure 3.6):

- **Trends by day (left pane):** The left pane of the display shows each of the user's measurements along with a *mean daily weight*. This mean daily weight is estimated from a model that accounts for several known biases in weight measurements,

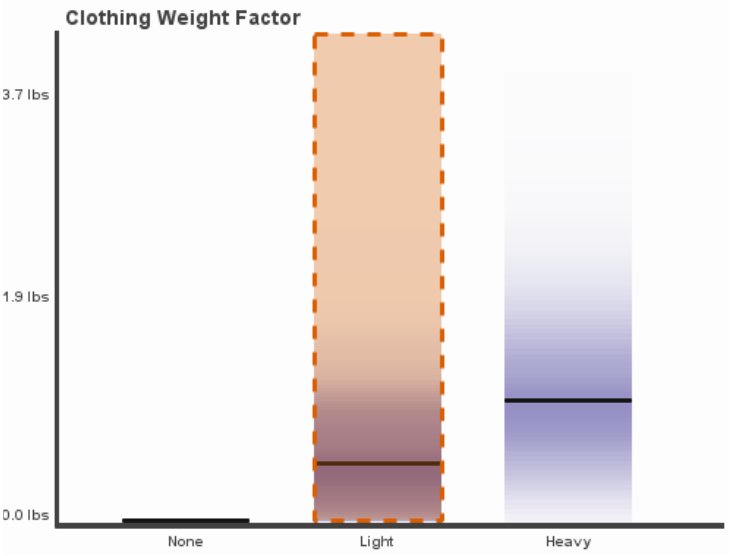


Figure 3.7. Enlarged view of top-right pane from Figure 3.6

thus presenting a consistent picture of a person's weight through *adopting an explicit model of weight*, described in more detail below.

- **Biases due to clothes (top right pane):** By asking people how much clothing they are wearing each time they measure (like in the intra-day weight study described earlier), we can adjust for biases in the weight measurement due to clothes, and present estimates of how much a person's clothes weigh to them.

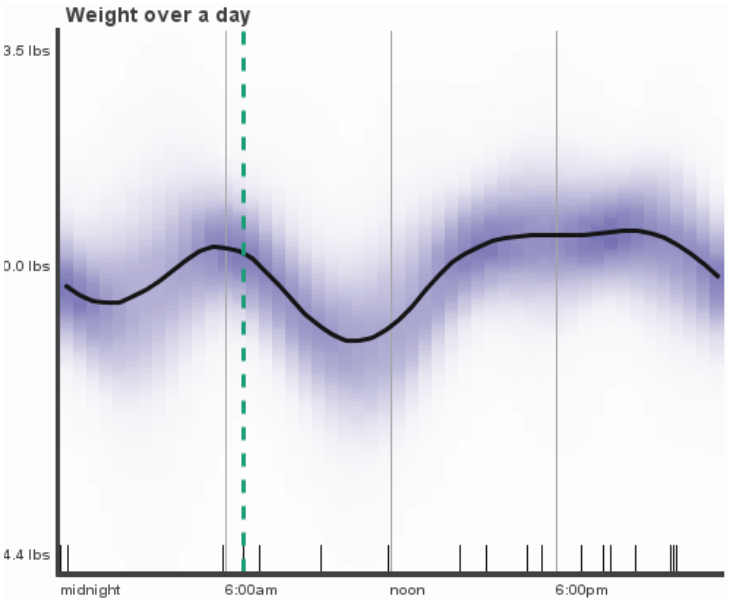


Figure 3.8. Enlarged view of bottom-right pane from Figure 3.6

- **Trends within the day (bottom right pane):** We can also adjust measurements based on what time of day a person weighs themselves, and show an estimate of this to users.

I use several principles in the design of the interface in order to help educate users about uncertainty. Natural language explanations describe two possible biases in a person's measurements (time of day and clothes), and brushing and color linking to connect these explanations to each graph: the natural language explanations appear for the most recent data point initially, and then for any point the user mouses over; then, the color of the text describing each bias (clothes in orange and time of day in green) is also used to highlight the corresponding effect in the right-hand panes. The uncertainty in estimates are plotted as gradient plots of posterior distributions from the model.

3.8.2 Weight model

When the user steps on the scale, a Bayesian autoregressive model is fit in order to estimate their day-over-day weight curve. This model includes a parameter for how much the person's weight can be expected to vary from one day to the next. It also includes a periodic submodel that uses a cosine fit with three harmonics to estimate a within-day curve describing how the user's weight fluctuates over the course of the day. Finally, the model includes effects to account for how much clothes the person is wearing (lightly or heavily clothed, with no clothes fixed at 0). The curve plotted in the left pane shows the person's mean daily weight assuming they are not wearing clothes. The model specification is as follows:

$$
\begin{aligned}
w_d^* &\sim \text{Normal}(w_{d-1}^*, \sigma^2) \\
w_{d,i} &\sim \text{StudentT}\left(w_d^* + Ll_i + Hh_i + \sum_{k=1}^{n}\left[\alpha_j \cos(2\pi k t_i) + \beta_j \sin(2\pi k t_i)\right], s^2, \nu\right) \\
l_i &= \begin{cases} 1 & if\ lightly\ clothed \\ 0 & otherwise \end{cases} \\
h_i &= \begin{cases} 1 & if\ heavily\ clothed \\ 0 & otherwise \end{cases} \\
t_i &\in [0,1) \quad (time\ of\ day)
\end{aligned}
$$

I first fit a similar model to the study in which I had people track their within-day weight. This model used hierarchical Normal parameters for the effects of clothes (l and h), the within-day weight curve (α and β), and day-over-day variance (s and ν). This allows me to set informed priors for these parameters for models fit to

individuals using the scale. Put another way, we have at least some prior belief for what the distribution of how heavy peoples clothes are, how their weight fluctuates within a day, and how much their weight fluctuates between days. The hope is that during deployment, a model like this will allow reasonable estimates to be made for individuals even when we do not already know much about them (but we know something about the population that they come from).

3.8.3 Deployment configuration

The scale also includes some features designed to aid in deployment: users of the scale can be randomly assigned to versions of the interface that have different feature sets, in order to test hypotheses about the value of different approaches to communicating uncertainty in this domain. For example, the interface might include (or not) the visualizations of posterior distributions (the gradient plots), or the natural language explanations of biases, or the plots of the biases themselves (the right-hand pane), or the interactive brushing and linking used to help participants explore the data. This will allow me to test (for example) whether the natural language explanations of uncertainty improve outcomes like trust or acceptability of error in the system.

3.9 CONCLUSION

The existing literature on intelligibility has used hypothetical surveys [61] to examine relationships between communicating uncertainty and acceptability of systems. This chapter extends those results by demonstrating evidence for decreased trust (and possible abandonment) of a simple, real-world sensing system due to a lack of communication of uncertainty. I found that people with a poor understanding of weight fluctuation trust their scales less, and that negative scale reviews are often based on a less sophisticated understanding of weight fluctuation. From my qualitative investigations, I derived design principles and a new weight scale design for communicating uncertainty more effectively, with the goal of educating users in order to increase trust and acceptability. This chapter asks if failing to communicate uncertainty can decrease trust in a real-world system, finding it does. Then, if we *should* communicate uncertainty, *how* can we do it in a way that people will understand? That is the subject of the next chapter.

4. Visualizing uncertainty in realtime bus arrival predictions

4.1 INTRODUCTION

Realtime bus arrival prediction is another domain where, as with weight scales, data is typically presented to users as single data points without any notion of uncertainty (e.g., see OneBusAway [104], which presents arrival predictions in the form of "*n* minutes from now" and "*n* minute delay", neither of which convey the expected error in those predictions). However, also as with weight scales, I believe that this missing uncertainty may have a strong impact on the usability of the system, and conveying it will substantially improve acceptability and trust, even without making the predictions more accurate.

For example, Susan might refer to a bus's predicted arrival time on a smartphone application (as in Figure 4.1) to check if she has time to get coffee before her bus to work arrives. She sees that the bus is running a few minutes late and is predicted to arrive in five minutes. There is no line at the coffee shop, so she steps in to order. However, the bus makes up lost time and arrives only two minutes later: Susan, still waiting for coffee, misses her bus and is late for a meeting.

Susan based her decision on a point estimate of arrival time, as presented in many predictive systems for bus arrival, flight time, or car travel. Her decision is reasonable given the point prediction she saw, but real-world predictions are subject to uncertainty (e.g., her bus is most likely to come in 5 minutes but may come in as little as 1 minute or as much as 9 minutes). Designers and analysts are responsible for reporting uncertainty with predictions to help people make decisions that align with their goals [16,90], yet most visualizations of predictions present the data as if it were true (Finger & Bizantz [26] as cited in Cook & Thomas [16]). Had Susan's application presented her with a more complete representation of the predicted arrival time—perhaps noting that arrival times earlier than 5 minutes are also quite probable—she may not have risked getting coffee.

Many attempts to communicate uncertainty rely on complex visual representations of probability distributions. For example, error bars and probability densities require prior experience with statistical models to correctly interpret [4,17]. People can better understand probabilistic information when it is framed in terms of discrete events. For instance, Hoffrage & Gigerenzer [37] found that more medical experts could accurately estimate the positive predictive value (precision) of a test when presented with discrete counts or outcomes. Discrete-event representations have been used to improve patient understanding of risk, e.g., by showing the uncertainty

This chapter is based on work published at *CHI 2016* with Tara Kola, Jessica Hullman, and Sean Munson [47].

This chapter centers on **RQ2:** *Can we build visualizations of uncertainty in continuous predictions (bus arrival times) that people do understand?*

For reference, **T1** is also repeated here:

By using richer probabilistic models of the uncertainty in predictive estimates—conveyed to users in ways that are relevant to their goals, that are in representations they can understand (e.g., by using frequencies and concrete examples instead of probabilities), and that are sensitive to the types of errors they care most about (e.g. false positives versus false negatives)—we can **A)** *improve users' acceptability of error in sensing and predictive systems,* **B)** *improve users' trust in those systems, and* **C)** *make it easier for users to answer questions they care about when using those systems.*

in a medical diagnosis as discrete possible outcomes (number of true positive, false positives, false negatives, and true negatives) [29]. However, these approaches tend to focus on binary prediction tasks instead of continuous predictions like time to arrival. In addition, the task of designing effective visualizations for realtime transit predictions is complicated by the context of use: people access these predictions on their mobile phones to make in-the-moment decisions that are time-constrained (providing little opportunity for training, interpretation, or complex interaction) using interfaces that are space-constrained (due to screen size). Existing discrete approaches to visualizing probability distributions typically requires a large amount of space or time to communicate the set of possible outcomes [38]. What might a compact, discrete-event visualization of a continuous probabilistic prediction look like?

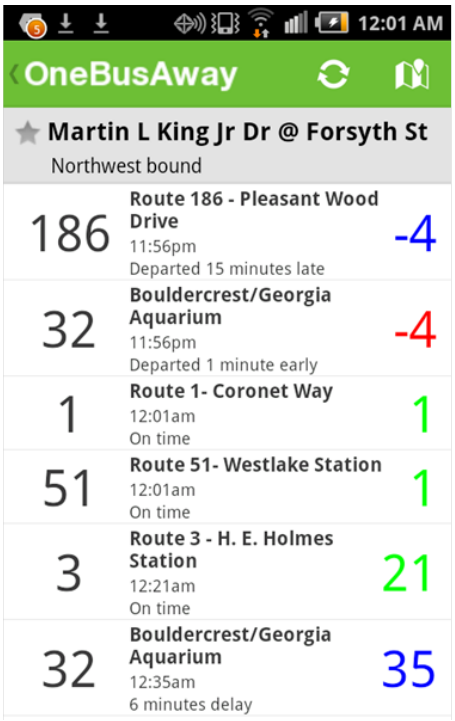


Figure 4.1. The existing OneBusAway [104] interface on Android, showing predicted time to arrival in minutes for various busses at a single bus stop.

I started by looking at user needs for communicating predictions and then designed and evaluated novel, goal-directed visualizations of hypothetical outcomes in a time- and space-constrained mobile application, realtime transit prediction. As a setting where users have direct, day-to-day experience with uncertainty, transit prediction provides a representative context in which to evaluate how well people can use different uncertainty visualizations: to answer **RQ2**, *Can we build visualizations of uncertainty in continuous predictions (bus arrival times) that people do understand?* More specifically, my goals were to see if and for what reasons people want uncertainty information in a bus setting; to identify effective uncertainty visualizations for realtime decision-making on a smartphone; and to test for differences in how precisely and confidently people can extract probabilities from different visualizations of uncertainty.

The three contributions of this chapter based on these goals are:

1) Develop **general and domain-specific design requirements and a rich description of user needs** for visualizing uncertainty in transit arrival times based on (i.) an analysis of the literature and (ii.) an initial survey of 172 people who use a popular realtime transit application.

2) Propose **design layouts and discrete-event visualizations of uncertainty** for conveying bus arrival time predictions on small screens based on an iterative design process. I introduce **quantile dotplots**, a novel modified dotplot that is a discrete analog to the common probability density plot.

3) Identify through a large user study of transit application users how accuracy and precision in estimating probability compares across several visualizations. Unlike previous work in communicating uncertainty in continuous outcomes [17,33], this is the first study to compare static discrete-outcome visualizations of probability distributions to continuous representations. I find that **a quantile dotplot depicting a small number of outcomes has ~1.15 times lower variance than a density plot**, making probability estimates 1-3 percentage points more precise.

My results further understanding of how to communicate prediction uncertainty to non-experts in everyday, mobile decision-making contexts. Specifically, I recommend using low-density dotplots (due to lower variance and higher user confidence) or density plots (which have only slightly higher variance, but were more visually appealing) for visualizing uncertainty in space-constrained environments.

4.2 DESIGN REQUIREMENTS FROM THE LITERATURE

To develop design guidelines for communicating uncertainty in realtime transit prediction, I will first establish baseline requirements for the effective communication of uncertainty based on the literature. **Each such requirement in this section is indicated in bold.**

4.2.1 Improving trust by communicating uncertainty

Research indicates that displaying uncertainty can improve trust and decision-making in everyday contexts. As I found in my studies of trust in body weight measurements in Chapter 3, single point estimates without uncertainty decrease trust and tend to be interpreted as being more precise than they actually are. In that Chapter I suggested **avoiding false precision in single point estimates** by displaying the uncertainty associated with weight data to improve trust. Similarly, Jung *et al.* [44] found that displaying the estimated remaining range of an electric vehicle as a gradient plot (i.e., with uncertainty) reduced range anxiety in a driving task compared to a single point estimate.

Joslyn & LeClerc [42] found that displaying uncertainty in weather predictions can lead to more optimal decision-making and trust in a forecast. When asked to make decisions about whether to salt roads (given a virtual budget, cost for salting, and cost for failing to salt when they should have), people made more optimal

decisions when given point estimates with probabilistic information. Subjects with access to probabilistic information even made more optimal decisions than subjects who were explicitly told the optimal decision based on a cot-benefit analysis. While the decision suggested by a cost-benefit analysis will give the best choice on average, always applying the decision will sometimes lead people to take precautions that seem unnecessary (e.g., salting the roads when the weather ultimately does not require it). After experiencing a such few errors, people may begin to distrust the strategy and ignore the suggested course of action. Probabilistic information, on the other hand, provides a more transparent form of information for decision making, leading to greater trust. I believe this insight also applies to realtime transit prediction: even if we could develop a system to make recommendations like "leave now to make your bus on time", **a more transparent communication of uncertainty** will maintain trust over the long term and leave people the agency to make mistakes.

4.2.2 Visualizing uncertainty

...1 As extrinsic annotation

A common approach to visualizing uncertainty is as an *extrinsic* annotation to a plot of the distribution's location (mean, median, or mode). For example, error bars representing confidence intervals or prediction intervals[1] can be superimposed on bar charts [4]. These intervals are extrinsic to other properties like the mean or mode since they are not integrated into the same encoding. By contrast, the probability density and mode are intrinsic to each other in a density plot, since the mode is visually encoded as the maximum of the density. Other distributional properties may be represented in summary plots using a series of marks (e.g., specific quantiles in a boxplot or modified box plot as in [17,39,74]).

Extrinsic representation can result in interpretation errors because the statistical construct (such as one standard error or a 95% confidence interval) is poorly understood [4,39], because individuals apply heuristics that are not correct (such as assuming that overlapping confidence intervals always indicate a non-significant difference [18]), or because the representation is ambiguous (such as an error bar being used to encode standard deviation, standard error, or 95% confidence interval). Finally, individuals tend to underweight probabilistic information (such as sample size or variance) when making judgments in favor of heuristic attributes like representativeness [94]. By separating the marks encoding underlying data from those encoding uncertainty, extrinsic representations are at risk of being viewed as peripheral, and consequently discounted when making judgments. Thus, to avoid ambiguity, simplify interpretation, and encourage users not to underweight probability information, I believe that **uncertainty should be intrinsic to the representation**.

1. In contrast to a *confidence interval*, which describes the precision of an inferred model parameter (e.g. population mean), a *prediction interval* is an interval that a given percentage of specific instances are predicted to fall into. While much of the literature focuses on confidence intervals (of interest to scientists using models for inference), we are more concerned with prediction intervals (of interest to an individual who wishes to know how likely their bus is to arrive in a specific instance).

Further, while a given prediction interval corresponds to a specific risk tolerance, a user may have differing risk thresholds in different contexts. For example, Susan may be willing to be late to her meeting 1/20 times, translating to a one-sided 95% prediction interval for estimated arrival time, yet she may tolerate more risk in different contexts like social gatherings or less important meetings. Different individuals are also likely to have different risk tolerances. Therefore, I believe that effective visualizations of uncertainty in this context should **allow users to apply situation-dependent risk tolerance**.

...2 As abstract, continuous outcomes

Many other abstract, static representations encode a predictive distribution's probability density function (PDF) intrinsically as *retinal variables* (e.g., color, shape, texture) [7]. For example, density plots encode the PDF as distance from the x-axis, violin plots encode it as width [3,45,87], and gradient plots encode it as opacity. Several studies that include variants of density and gradient plots find little evidence of a performance difference between the two [17,39]. Opacity is a less effective encoding than height, width, or area [63]. As a result, I do not test the gradient plot.

Not all encodings of continuous outcomes using retinal variables make distributional properties intrinsic. Ibrekk & Morgan [39] compare density plots to plots of cumulative density functions (CDFs), amongst several other encodings. CDFs encode cumulative density as distance from the x-axis, allowing the probability of intervals to be estimated from height. They found that CDFs were unfamiliar to participants and required training. Not surprisingly, people had particular difficulty in using CDFs to estimate means, most likely because there is no simple visual variable that corresponds to mean (nor mode) on a CDF.

...3 As hypothetical, discrete outcomes

I use *discrete outcomes* to refer to techniques that employ draws from a probability distribution rather than abstract probabilities of events. Discrete approaches were initially found to improve reasoning in textual communication. Gigerenzer and Hoffrage [30] found that statistical word problems described in terms of *natural frequencies* (e.g., 10/100) rather than probabilities (10%) were more likely to elicit inferences according to Bayes' rule in laypeople. Past work on visualizing uncertainty through hypothetical, discrete outcomes uses spatial or temporal bandwidth to communicate. For example, Garcia-Retamero and Cokely [29] reviewed studies of several types of visual aids for communicating health risks, including discrete outcome charts that illustrate treatment risk: they found that displaying *icon arrays* (a grid of pictograms, each representing a patient who lived or died) improved the accuracy of people's risk assessment. Hullman *et al.* use animation to display

discrete outcomes more compactly in space [38], finding that animated discrete outcomes (called *hypothetical outcome plots*, or HOPS) support more accurate probability estimates than static alternatives (violin plots and error bars) for some tasks. However, by presenting outcomes over time, animated techniques bring a time-precision trade-off: to make more precise inferences, a user must view more outcomes, taking more time [ibid].

From the evidence in both visual communication and statistical reasoning, I believe that **discrete outcomes can improve decision making under uncertainty**. However, because transit decisions are often made quickly in real time I focused on developing non-animated presentations of discrete outcomes that are **glanceable** yet compact enough for a mobile phone display, in which it is typical to visualize the upcoming arrival of ~10 buses on one screen [25].

4.2.3 Visualization in space-constrained environments

To display many buses simultaneously on a mobile phone screen, visualizations should be **compact**. Techniques like horizon graphs [36] and sparklines [93] have been proposed for visualizing time-series data in space-constrained environments. Visualizing uncertainty in transit arrival predictions encounters similar issues as these approaches; for example, a probability density function of predicted arrival time will become quite tall as its variance decreases (particularly, close to the predicted arrival time the prediction will become very precise). My work demonstrates possible solutions for the specific context of visualizing PDFs on mobile phone displays.

4.3 SURVEY OF EXISTING USERS

While this reading of the literature provides an initial grounding for design, to apply these results to a user-centered uncertainty visualization also requires an understanding of user goals. To establish design criteria for representations of uncertainty based on user needs, I surveyed users of one popular realtime transit application, OneBusAway [25].

4.3.1 Method

I conducted a survey to identify 1) how users currently use realtime bus arrival predictions and 2) their unaddressed needs for goal-oriented uncertainty information. I surveyed 172 users of OneBusAway, recruited via social media and department mailing lists.

...1 Users' existing goals

To identify important user scenarios to address and what types of information are most important to those scenarios, I asked people about the *primary goals*

they have when using OneBusAway. I developed a set of possible questions (e.g., "When should I start walking to the bus stop to catch my bus?"[2]) that people may ask using the interface using observations from previous studies of OneBusAway [43], my own reflections on using the system, from informal interviews with a small group (~15) of other users at my university, and through piloting the survey. I presented participants with a list of 9 such questions, and asked how often (on a 7-point scale from "never" to "always") they try to answer each question using OneBusAway. I also asked them if there are other ways they use OneBusAway in an open-ended question.

2. The full survey is available in the supplementary material of [47].

...2 Problems with OneBusAway and unaddressed needs

I similarly presented participants with a list of types of information not currently provided by OneBusAway and asked them to rate the *potential helpfulness* of these (on a 5 point scale from "not helpful at all" to "very helpful"). I also provided an open-ended question asking about needs for uncertainty information not in this list. Finally, I asked people to describe the *worst experience* they have had using OneBusAway's predictions.

4.3.2 Results and Discussion

...1 Users' existing goals

The top 5 highest-rated questions users currently ask are:

- **When to leave:** When should I start walking to my bus?
- **Wait time:** If I leave now, how long will I have to wait at the bus stop?
- **Time to next bus:** I missed my bus, how long will I have to wait for the next one to come?
- **Schedule risk:** Will I get to a meeting/event on time despite bus delays? This relates to a commonly-described worst experience of buses coming later than expected. For example:

  A more recent bad experience was when I was waiting for the 511 or 512 for over an hour. At least five buses should have passed, but they either did not show up or they were full and didn't let anyone on

- **Schedule opportunity:** Will I have enough time to do ____ before the bus arrives? This relates to a commonly-described worst experience of the bus coming

earlier than expected after someone has used the prediction to decide to do something else before going to the bus; e.g.:

It showed delays on a bus due to which I didn't leave home as I didn't want to wait at the bus stop for long (the bus stop is 4 mins from my home), but it suddenly came on time and I missed it. Sometimes, it even comes early when it shows delay.

...2 Problems with OneBusAway and unaddressed needs

The top three questions users would like to be able to ask, but which are not well-supported by the current OneBusAway interface, are:

- **Status probability:** What is the chance OBA is showing the correct arrival status? This problem was also reflected in a commonly-described worst experience, wherein the bus never shows up and people have to make alternative plans. For example:

  My bus is perpetually 9 minutes away...while I watch alternative buses pass me thinking that oh, mine is going to be here soon only to eventually see "no information" for my bus. I could have been on my backup bus a half hour ago!!!

  It was common for people to report their worst experiences were related to status probability: for example, OneBusAway said "departed", but the bus had not arrived; it said "arriving" but had already departed. Any noisy estimate reduced to a categorical status will exhibit these types of errors which could be mitigated by conveying status probabilistically.

- **Prediction variance:** What is the chance the predicted arrival time will change unexpectedly?

- **Schedule frequency:** How frequently do buses arrive at various times in the day?

4.4 DESIGN REQUIREMENTS

Based on my literature review and user survey, I identified the following necessary design elements:

Point estimate of time to arrival: To support *glanceability*, the point estimate of arrival time is necessary: people often use OneBusAway to make fast decisions about when to arrive at the bus stop. In addition, previous work has found that even

when providing probabilistic estimates, people still want a point estimate. The existing point estimate of OneBusAway supports estimation tasks from my survey like *when to leave*, *wait time*, and *time to next bus*, though without communicating risk.

Probabilistic estimate of time to arrival: While people often want a point estimate of arrival time, a point estimate without uncertainty will often convey a *false precision*. A probabilistic estimate will help users understand that there is a chance the bus will come earlier or later than the point estimate. This helps people assess *schedule risk* and *schedule opportunities*. A probabilistic estimate also allows people to make conservative estimates while planning for meetings, or less conservative estimates for low risk situations – that is probabilistic estimates *allow situation-dependent risk tolerance*. This will help people better answer questions about *when to leave*, *wait time*, and *time to next bus* (the highest rated goals) and prepare people for commonly-reported *worst experiences* like a bus coming unexpectedly early or late.

Probabilistic estimate of arrival status: For example, what is the chance the bus has already arrived? Among questions not currently supported by OneBusAway, survey respondents most wanted support for this question (*status probability*), and commonly reported *worst experiences* related to it.

Data freshness: Because OneBusAway does not currently give probabilistic estimates, one of the only available signals for expert users to assess risk is *data freshness*: OneBusAway indicates the time of the last update for realtime predictions and whether the current prediction is based on realtime data (it reflects the scheduled arrival time when realtime data is not available). This freshness information should either be provided to users in a redesigned interface, or should be incorporated into any models driving probabilistic estimates.

I believe these design elements will address each goal identified in the user survey with the exception of the goal of knowing *schedule frequency*. I felt that this goal is better addressed through a separate interface, such as a trip planner or schedule explorer in a mapping application. Schedule frequency is less relevant to in-the-moment decision-making than it is to long-term planning (can I rely on a bus arriving within some amount of time?). When schedule frequency is relevant to in-the-moment decisions, it typically reduces to other goals, like *time to next bus*.

4.5 DESIGN

I conducted an iterative design process focused on the design requirements set out above. This process began with a wide exploration of ideas through sketching, followed by paper prototyping in increasing fidelity, and culminated in digital mockups. These phases were informed by ongoing user feedback gained through informal down-the-hall testing with a total of 24 users. During informal testing,

Figure 4.2. Alternative layouts we developed.

A) Bus Timeline: Each row (timeline) shows one predicted bus.

B) Route Timeline: Each row shows all predicted buses from a given route.

users were presented with hypothetical scenarios of use and asked them to think aloud as they interpreted the display.

Many of the design issues encountered are somewhat orthogonal to specific of encodings of probability: given a particular timeline layout, for example, we could encode probability in many ways (e.g., as area, discrete events, a gradient). I first present my proposed set of designs and their rationale, then discuss possible techniques for encoding probability on small screens.

4.5.1 Proposed designs and rationale

My proposed designs, instantiated with one particular visualization of uncertainty (density plot) out of several possible, are shown in Figure 4.2. Several design decisions were necessary to resolve design tensions and to match user goals.

...1 Different layouts better serve different use cases

I developed two alternative layouts, *bus-timeline* and *route-timeline*. The *bus-timeline* layout gives a timeline for a single bus on each row, similar to how the existing OneBusAway app displays a single row per bus, sorted by predicted time to arrival. This simplifies understanding and navigation, but is less compact in addressing problems like assessing *schedule frequency*, and, once the probabilistic visualizations are added, less compact than the current application. *Route-timeline*, by contrast, creates

a more complex display and navigation (requiring navigation in two dimensions), but more easily aids understanding of *schedule frequency* (how often is the bus) and *schedule opportunity* (since if one is considering the risk associated with missing the next bus, it is easier to see how soon the bus after that is coming and factor that into one's decision).

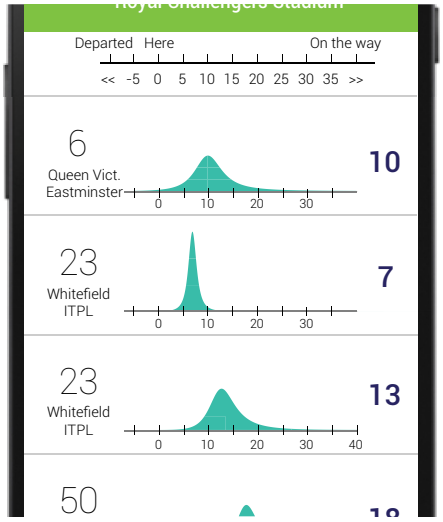


Figure 4.3. An example of a design we rejected for placing point predictions (along the right side) outside the context of uncertainty, making it more likely to give users a false sense of precision.

...2 Point estimates and probabilistic estimates should coincide spatially

I explored several tradeoffs between prominent point estimates versus probabilistic estimates, what I call the **glanceability/false precision tradeoff**. A too-prominent display of the point estimate causes users to ignore the probabilistic one, thus still giving a false sense of precision; a less-glanceable point estimate will be difficult to skim and frustrating to use. We want a display that is *glanceable* but which also does not convey false precision. To resolve this, I concluded that these two elements should coincide spatially: that is, *looking at the point estimate should encourage the user to also be looking at the probabilistic estimate*. I had considered designs in which the point estimate was along the right-hand edge of the display (Figure 4.3), as in the original OneBusAway. I concluded that this facilitated glanceability, but also allowed users to pay too little attention to the probabilistic estimates. Moving the point estimate onto the probability distribution resolved this tension.

...3 Annotated timelines give probabilistic estimates of status "for free"

While I considered designs that more explicitly communicate the probability that the bus has arrived, I realized that an annotated timeline combined with probabilistic predictions communicates this implicitly. By denoting areas that correspond to "departed", "now", and "on the way" on the timeline, users can directly read these probabilities from the distributions depicted; see the timeline annotations across the top of Figure 4.2.

...4 When to leave is implicit in time to arrival

I considered designs that communicated when someone should leave to catch their bus; i.e. designs that directly addressed the *when to leave* goal. However, there are several difficulties with this approach: first, when to leave is not the only goal for which people use OneBusAway; thus it would need to be integrated into displays

communicating information like time to arrival (or alternate designs developed for both goals). This exacerbates space issues. Estimating when to leave also requires substantial knowledge about the users' plans, and introduces further uncertainty (e.g., how long does it take to walk to the stop?).

...5 Data freshness may be subsumed by an improved model

OneBusAway often does not have truly realtime information, but instead updates when buses check in. As noted previously, expert users often refer to the last check-in time as a way to evaluate how much they trust the application's prediction. To facilitate this use, I considered several designs that included indicators of data freshness or last update times. Ultimately I decided not to include this information, as the model used to generate the probabilistic arrival information should **take data freshness into account to provide better estimates to all users**, rather than continuing to support a workaround used by expert users.

...6 Synchronized timelines allow comparison between buses

In my designs, the axis of the timeline in each row is synchronized to the other rows, facilitating comparison between buses. I considered designs with each row having its own time range depending on the prediction (e.g., one row with low variance might show a density plot covering 5-10 minutes from now; another with high variance might have an axis covering 5-15 minutes from now). However, such relative timelines are very difficult to compare between buses on different rows—buses with different variance might look similar because the relative timeline would also cause the density to be scaled.

4.5.2 Encoding probability in space-constrained environments

Given the chosen design, we need an effective way to encode probability at small sizes. I considered several approaches (Figure 4.5). Most of these are drawn from the literature, including density plots, violin plots, and gradient plots. I also propose variants of two existing discrete plots for visualizing predictive distributions as discrete outcomes, stripeplots and dotplots.

...1 Discrete outcome visualizations of continuous variables

I explored several ways to convey a continuous predictive probability distribution as discrete outcomes. The first is based on Wilkinson's dotplots [105], which are typically used to communicate the distribution of experimental samples (e.g., [68]). I instead adopt these plots to display theoretical quantiles from a predictive distribution. As Wilkinson notes, correctly-produced dotplots have the desirable property of also conveying the density of the distribution. My **quantile dotplots**

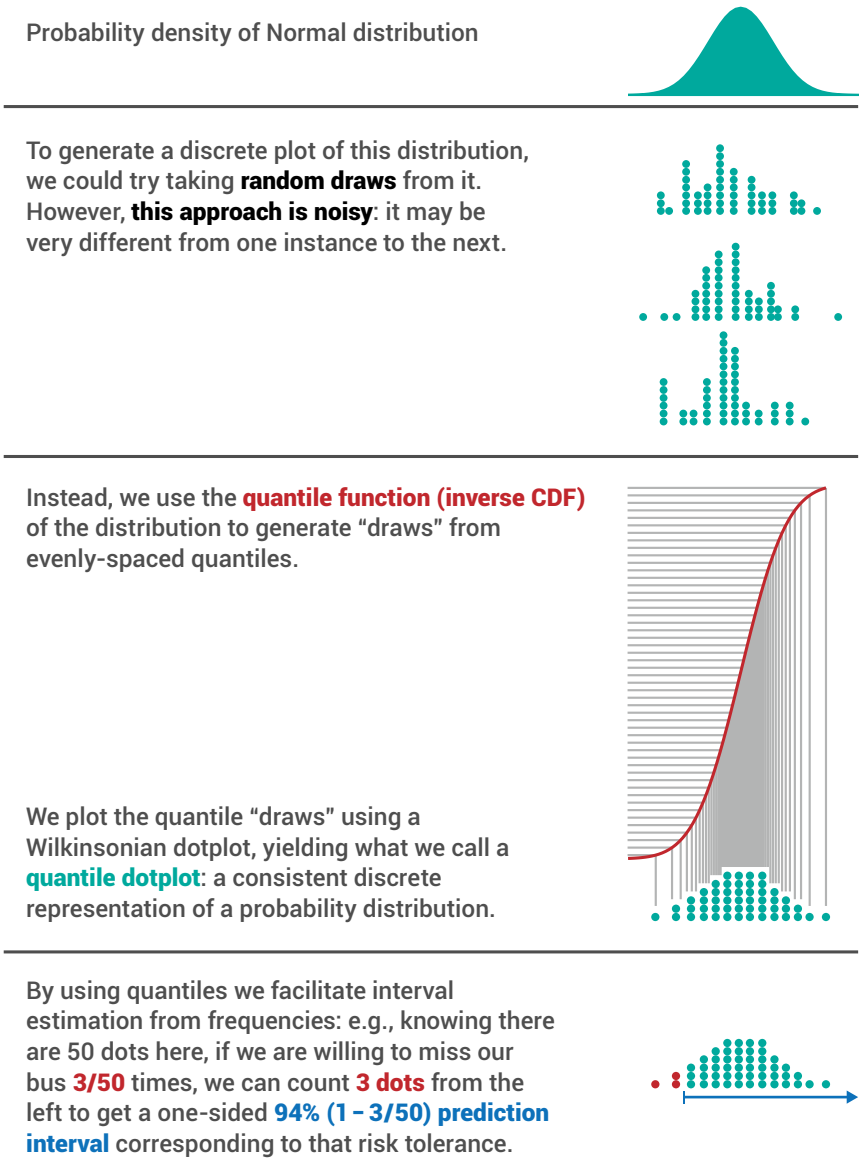


Figure 4.4. Explanation of quantile dotplots.

have this property, as well as the additional property of allowing direct estimation of arbitrary (to a certain precision) predictive intervals through counting (see Figure 4.4). I believe that this form of natural reasoning about predictive intervals—as frequencies—should allow people to obtain precise estimates of predictive intervals in a way that is easily understood.

I also use **stripeplots** [24] of theoretical quantiles to communicate a continuous probabilistic prediction as hypothetical outcomes. In these, the density of stripes in a region encodes probability density, and as in quantile dotplots (though less easily), predictive intervals can be estimated directly through counting. Where dotplots are a discrete analog to a density plot, stripeplots can be thought of as the discrete analog to a gradient plot.

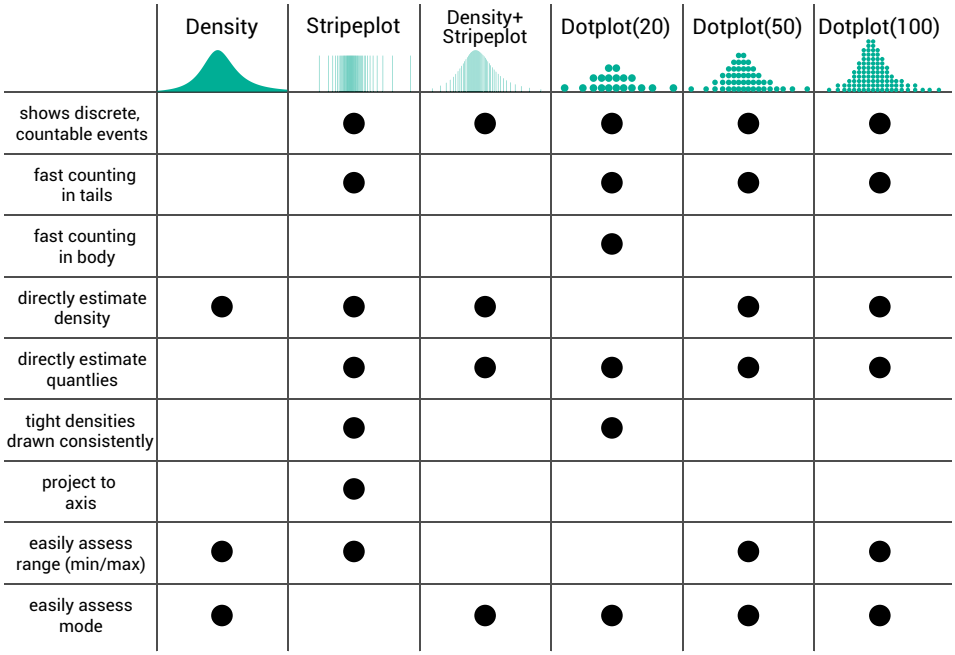


Figure 4.5. Comparison of various encodings of probability we considered for use in the designs.

...2 Tight densities require special attention on small screens

Displaying many rows of predictions on a small screen necessitates relatively small row height. Unfortunately, distributions with low variance will become very tall, exceeding the row height. Traditional solutions include horizon charts [36] (which I suspect are unfamiliar to lay users), or normalizing all density plots to the same height (which makes comparison difficult). This problem is most pronounced on buses with tight variance, i.e., the most precise predictions. Consequently, for density plots I adopted the compromise approach of scaling down the max height only when it exceeds the row height. This adjustment affects only the predictions of which the model is most certain, so fine-grained resolution of probability becomes less important to most goals. This adjustment is required only for *density*, *dotplot-50*, and *dotplot-100* (in the dense dotplots, instead of scaling I reduce the dot-spacing). Dotplot-20 and stripeplot have the advantage of a *consistent representation of probability in tight densities*: they need not be modified.

...3 Countability may vary from tails to body

Care must be taken in deciding how many hypothetical *draws* (quantiles) to include in discrete plots. Figure 4.5 compares some of the tradeoffs here: With few draws, as in *dotplot-20*, it is easy to count the dots in the tails and body of the distribution, but the density is less well-resolved. With many dots, as in dotplot-100, counting in the tails is often still easy, but in the body overwhelming; however, density is very well-resolved.

...4 Selected encodings

To select the encodings to evaluate for my final design, I constructed the matrix shown in Figure 4.5 comparing various properties of the encodings. I selected *density*, *stripeplot-50*, *dotplot-20*, and *dotplot-100* as representing a wide range of possible trade-offs suggested by this matrix.

4.6 EXPERIMENT

I conducted an online survey to evaluate the effectiveness of my designs in conveying uncertainty. The goal of this survey was to assess how well people can interpret probabilistic predictions from the visualizations and to elicit their preferences for how the data should be displayed.

4.6.1 Method

To assess how well people can judge probability from the visualizations, I adopted an approach similar to that of Ibrekk and Morgan [39], who presented various representations of uncertainty for weather forecasts and asked subjects to report probabilities (e.g., snowfall >2 inches, or between 2 and 12 inches).

I created four scenarios based on the goals identified in the user survey, each with two questions about the probability of bus arrival. For example, in one scenario the respondent is waiting for a bus, and must decide if they have enough time to get coffee before the bus arrives. They are asked what the chance is that the bus will arrive 10 minutes or earlier, and respond using a visual analog scale, a 100-point slider from 0/100 to 100/100. I call their response the *estimated p* (in contrast to the

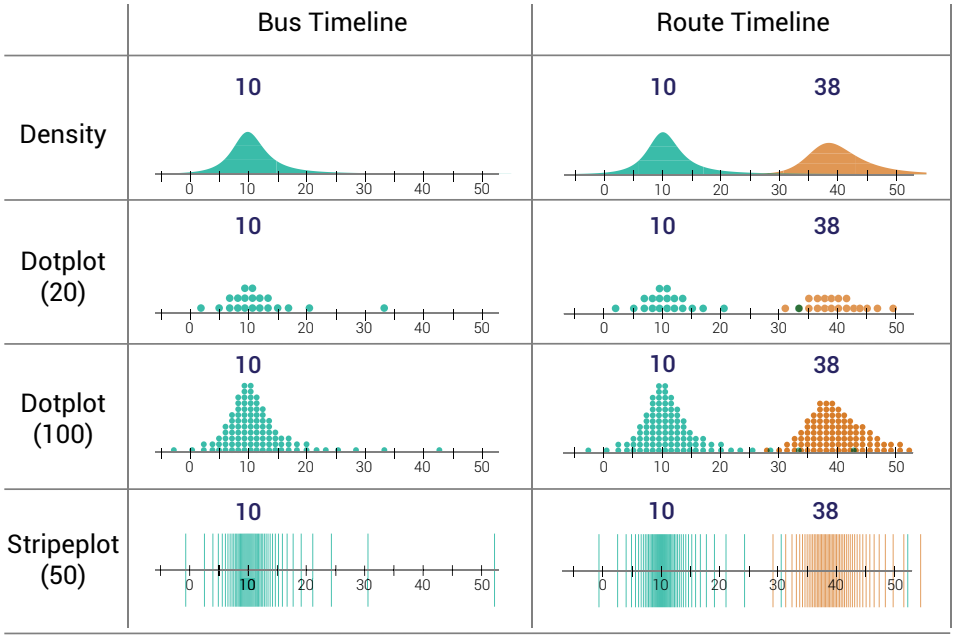


Figure 4.6. The four types of visualizations selected for evaluation.

true p, which I calculate from the underlying probability distribution). A bubble on the response slider shows this chance expressed in all three denominators used by the various visualization types (e.g. "20/100, 10/50, 4/20"), so that participants do not have to do mental arithmetic in the dotplot and stripeplot conditions. The predictions in each scenario were generated from models based on Box-Cox *t* distributions [78] fit to ~2 weeks worth of arrival time data for actual buses in Seattle, but the buses were given fake route names. Participants are also asked how *confident* they are in each probability they estimate. At the end of the survey they rate the *ease of use* and *visual appeal* of each visualization. All subjective ratings are made on 100-point visual analog scales.

Scenario order was randomized between participants. Each participant saw each *visualization type* (*density*, *stripeplot*, *dotplot-20*, or *dotplot-100*) once. Before each scenario, they were also given a brief tutorial explaining the encoding they were about to use. Pairings between scenario and visualization type were also randomized. Participants were also randomly assigned to see all visualizations in the *bus-timeline* or *route-timeline layout*. A full version of the survey can be found in the supplementary material of [47].

...1 Participants

I recruited participants from a variety of locations, including department mailing lists, a local transit blog, and a local forum on reddit.com. Participants were entered into a raffle for 1 $100 Amazon.com gift card and an additional $25 gift card per 100 participants. Since my primary research questions were about the effect of visualization types, not layout, I ran the first 100 participants only on the bus-timeline condition. This threshold was chosen based on a power analysis of data from Hullman *et al.* [38], which suggested a power of at least .8 for detecting similar effect sizes to that study after 100 participants. After reaching 100 participants in the *bus-timeline* layout, the remainder of participants were randomly assigned to either the *bus-timeline* or *route-timeline* layout. After removing 9 participants for incomplete data, I had 320 participants in the *bus-timeline* and 221 participants in the *route-timeline* layouts. Participants skewed male (71% male). 90% were existing OneBusAway users.

4.6.2 Results

To understand how well each visualization performs, we can examine the error in people's probability estimates. I break *error* into *bias* (do people over- or underestimate probabilities on average?) and *variance* (how self-consistent are people's estimates, whether biased or not?). **So long as the bias is low, I believe that variance is the more important component of error in this task**: over time people

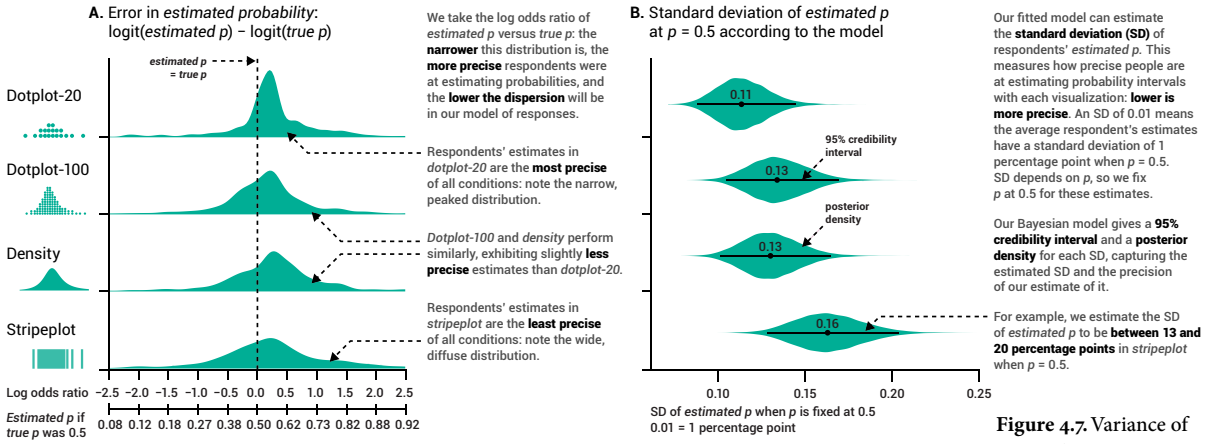


Figure 4.7. Variance of respondent estimates of probability intervals, **A)** as raw data and **B)** as estimated by the model.

can adjust their risk tolerance to a small but consistent bias, but they cannot do so if their estimates are not consistent. We will consider overall error, bias, and variance in turn.

...1 Overall error in participants' probability estimates

We can start by looking at the overall shape of participants' estimation error: logit(*estimated p*) – logit(*true p*) for each question.[3] Figure 4.7A shows the density of those differences, broken down by visualization type. The bias in responses is consistently low and positive across conditions: note that the error distributions all peak in approximately the same place, slightly to the right of 0 (the dashed line). *Variance* appears to be lower in *dotplot-20* compared to the other visualizations: the distribution of error is narrower. However, this does not necessarily equate to more *self-consistent* responses within a participant. I therefore use a model to assess bias and variance more systematically and to account for within participant effects.

3. The logit function is an s-shaped function that transforms probabilities into log-odds, often used when to simplify the analysis of probabilities by transforming them onto the unbounded real line.

...2 Regression model for bias and variance

I fit a beta regression to participants' estimated probabilities, which assumes responses are distributed according to a beta distribution. This distribution is defined on (0, 1) and naturally accounts for the fact that responses on a bounded interval have non-constant variance.[4] In other words, the variance of *estimated p* changes with the probability being estimated. For example, at *probability* = 0.5 one can guess 0.5 + 0.4 = 0.9; at *probability* = 0.9 one cannot guess 0.9 + 0.4 = 1.3 (it is greater than 1.0), so responses "bunch up" [86] under 1.0 and variance is lower. Beta regression has been shown to be better-suited to this type of data than linear regression [86].

My regression uses a submodel for the mean (in logit-space) and the dispersion (proportional to variance, in log-space) [86]. This allows us to model the bias of people's *estimated p* as effects on the mean of their responses, and the variance as effects on the dispersion of their responses. Specifically, I include *visualization*,

4. Because 0 and 1 are not defined in the beta distribution, we treat answers of 0 and 1 from the visual analog scales as 0.001 and 0.999.

logit(*true p*), and their interaction as fixed effects on mean response. I include *visualization*, *layout*, and *gender* as fixed effects on the dispersion (in other words, some visualizations or layouts may be harder to use, resulting in more variable responses; and men may be better or worse at this task). I also include *participant* and *participant* × *visualization* as random effects (some people may be worse at this task, or worse at this task on specific visualizations), and *question* as a random effect (some questions may be harder).

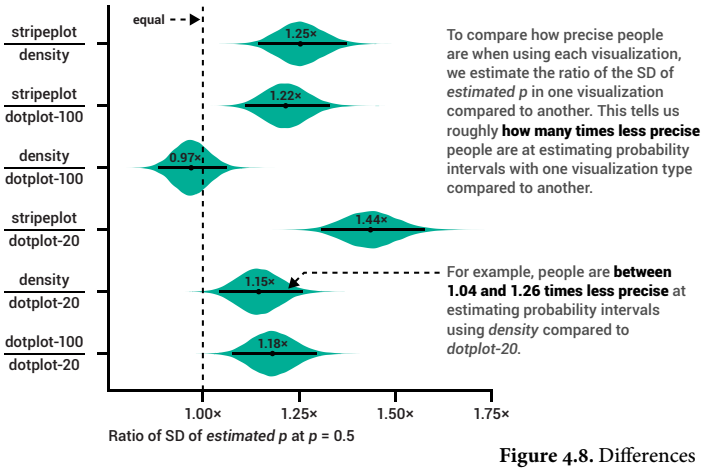


Figure 4.8. Differences in variance for each visualization type.

I use a Bayesian model, which allows us to build on previous results by specifying prior information for effects, and report results primarily as posterior distributions with 95% credibility intervals (the Bayesian analog to a confidence interval) [55,56]. I derive priors from fitting a similar model to the data from Hullman *et al.*[38], which had a similar task (estimating cumulative probabilities on three visualizations: a violin plot, animated hypothetical outcomes, and error bars). I set Gaussian priors for fixed effects in my model that capture the sizes of effects seen in the Hullman *et al.* data within 1-2 standard deviations, with skeptical means (0 for intercept and 1 for slope in logit-logit space, corresponding to an unbiased observer). I use the posterior estimate of the variance of the random effect of participant in that model as the prior for the variance of random effects in my analysis. Full priors and posterior estimates are available with my data.[5]

...3 Bias in respondent probability estimates

Consistent with Figure 4.7A, my regression found that estimates were slightly biased on average, and these biases were similar across conditions (more details in supplementary material of [47]). My beta regression model accounts for this bias when estimating the variance of participant responses. The slight overestimation here may be because all of the distributions are right-tailed (positively skewed). This is generally true of transit arrival time data; thus, if the skew is the source of this bias we should expect to see this effect in real-world situations in this domain but perhaps not others. Skewness of distributions is known to affect risk aversion in financial decisions made from density plots [100]; these biases may be related.

5. Similar results were obtained using more default priors, so the results are not highly sensitive to choice of priors. The model was fit using Stan [88], with 16 chains having 20,000 iterations each (half warmup), thinned at 8, for a final sample size of 20,000. Parameters of interest all had effective sample sizes > 10,000 and potential scale reduction factor < 1.001. See https://github.com/mjskay/when-ish-is-my-bus for survey data and code (DOI: 10.6084/m9.figshare.2061876)

...4 Variance in participant probability estimates

As noted above, I believe that variance is more important than bias in this task, as low variance would allow people to adjust their behavior to a consistent bias over long-term usage. We can estimate the variance associated with each visualization as a standard deviation in *estimated p* if *p* is fixed at 0.5 (Figure 4.7B). Figure 4.8 shows pairwise comparisons of SD for all visualizations (pair specified in left column). *Dotplot-20* has the lowest estimated variance (SD of ~11 percentage points), being about 1.15 times more precise than density plots. By contrast, *dotplot-100* has similar variance to *density*, consistent with people estimating area instead of counting dots, perhaps because there are more dots than they may be willing to count.

...5 Confidence

Ideally, greater confidence in a given answer would be associated with less error, indicating that people are able to self-assess their accuracy. I used a similar beta regression to model *confidence* in estimates depending on *visualization*. Participants expressed higher confidence in their estimates on average in the *dotplot-20* condition (mean = 81/100, 95% CI: [77, 83]) than the next-most-confident condition, *dotplot-100* (mean = 73, 95% CI: [71, 76]). At the same time, confidence in the *dotplot-20* condition correlated negatively with absolute estimation error (Spearman's $\rho = -0.18$, 95% CI: [−0.13, −0.25]), an association I did not see in other conditions. At least with *dotplot-20*, people have some ability to assess how good their own estimates are. I suspect that this may be due to the fact that with *dotplot-20* one can choose either to be precise (by counting dots) or to give a less precise, less confident answer (by approximating density or area instead of counting).

...6 Ease of use and visual appeal

I also analyzed ease of use and visual appeal using beta regression. *Density* had the highest *visual appeal* (mean = 66, 95% CI: [64, 67]); *dotplot-20* was less visually appealing (mean = 43, 95% CI: [42, 45]). However, despite these differences, ease of use for all visualizations except *stripeplot* was ~60 (stripeplot mean = 35, 95% CI: [33, 36]), suggesting only *stripeplot* was found consistently difficult to use. This may reflect *stripeplot*'s much higher estimation variance than the other visualizations: it had a higher standard deviation by about 4-5 percentage points when probability = 0.5 (Figure 4.7B), or about 1.44 times the SD of *dotplot-20* (Figure 4.8).

4.7 DISCUSSION

4.7.1 Discrete outcomes work best in small numbers

My results suggest that discrete-outcome visualizations of uncertainty can improve probability estimation in space-constrained visualizations of continuous outcomes

if care is taken in their instantiation. While *dotplot-20* improved estimation variance over *density*, *dotplot-100* performed very similarly to *density*. In addition, *Stripeplot* performed very poorly. I believe this may reflect the principle that *discrete plots with too many outcomes converge to continuous encodings*: since counting dots is arduous in dotplot-100 and stripeplot-50, people are more likely to read them like density plots and gradient plots (respectively), nullifying the value of the discrete outcomes. In dotplot-20, people can count quickly by using subitizing, the ability to quickly recognize rather than count groups of items in number < ~5 [12,31]. Since the vertical groups of dots in dotplot-20 are rarely are over 5 dots high, interval judgements (particularly close to the tails) are often reduced to quick, accurate judgments through subitizing. Thus, **I recommend discrete outcome plots with few enough outcomes to take advantage of subitizing.**

4.7.2 Implications for design and future work

...1 The value of communicating uncertainty

In the first survey, users described goals and unfortunate experiences in OneBusAway that information about uncertainty could help mitigate. In the second survey, most respondents said they appreciated the idea of representing uncertainty. They said this information could help them make better decisions, alleviate their anxiety when the app's information does not match their knowledge, or help them with a problem they commonly experience with OneBusAway.

A minority of respondents said they did not care about the uncertainty information: that point estimates are sufficient. In contrast to respondents who said the information could help, these respondents tended to say the point prediction presented in OneBusAway was consistently accurate. An additional minority actively did not want to receive any information about uncertainty. Several of these people compared evaluating probability information in visualizations to statistics courses. Five feared that, if given uncertainty information, they would become responsible for making decisions, and would have to take responsibility for the wrong decision: "you're more likely to be unhappy than if you missed the bus and can just blame the app." While this represents a small number of participants, I believe that future work is necessary to see how widespread such reactions may be in real-world deployments.

...2 Navigating the precision versus glanceability tradeoff

Designers should attend to the balance of precision and glanceability in representing uncertainty. Participants were divided over whether the visualizations were appropriately glanceable for a transit mobile app. While some said the new designs were easy and clear, more described feeling overwhelmed at least in the context of the

experiment and the majority of commenters expressed doubts about being able to use these visualizations while walking to a bus stop. Despite respondent concerns about whether they or others could understand the visualizations, the survey results overwhelmingly show that people understood them.

The designs presented here should be evaluated in longitudinal field studies to assess actual acceptability and use. For example, survey respondents were concerned that the dot plots would compel them to count, but in practice they may find that they count when they want precise estimates but are able to get a good overview from a quick glance. As transit prediction is an everyday practice for many, more sophisticated use of the visualizations may develop over time, making learning an important component to understand in this space.

The designs I evaluated also did not fully exploit interactivity, which might enhance the glanceability of the current static visualizations while preserving uncertainty information. For example, I prototyped a "risk slider" that lets people move the point estimate to match a specific risk threshold; this would allow them to fit the point estimate to their overall preferences or change it to match a specific situation. This feature can be incorporated into any of the proposed designs and should be evaluated as a technique to help resolve the glanceability/false precision tradeoff.

My research demonstrates that the visualizations can help people accurately, precisely, and confidently evaluate uncertainty, laying the groundwork for future studies evaluating the effects of differences in precision on behavioral measures.

...3 Visual appeal vs. estimation tradeoff

Related to the precision/glanceability tradeoff, people were also divided about preferring the dot plots or the density plots. The dotplots, while ~1.15 times more precise than the density plots and yielding higher confidence, were also rated less visually appealing. I do not know if this is a consequence of unfamiliarity, or if it is because the dotplots are visually busier. It is worth investigating whether the improvement from dotplots is worth decreased visual appeal, or if participants might get used to the dotplots over time.

4.8 CONCLUSION

Building on my intuition from Chapter 3—that the usability of realtime transit arrival predictions, like weight scales, could also be improved through communicating uncertainty—in this chapter I set out to answer the question, just *how* do we communicate uncertainty in a way people can actually understand? I investigated this question in the context of an existing transit prediction application, OneBusAway [104], first studying existing users' needs for probabilistic prediction. I also drew upon literature in statistical reasoning that suggests discrete-event visualizations

of uncertainty elicit better probabilistic estimates from people [30]. There being no existing discrete-event visualization of continuous probabilistic predictions suitable to the time- and space-constrained context of a transit prediction app, I developed a novel discrete visualization of a continuous probability distribution I call the *quantile dotplot*. When used with a small number of draws (say 20), this visualization technique combines prior results in statistical reasoning [30] with the human capability of subitizing [12,31] to improve the precision of people's probabilistic estimates by about 15%.

Both this and the previous chapter have focused on the communication of uncertainty to the end-user—the effects of poor communication, and a possible approach to communicating uncertainty more effectively when we do. However, I believe that communicating uncertainty is most effective when considered throughout the entire stack of a sensing or predictive system. The models we choose effect what we can communicate, and the expectations of our users effect the models we should choose. That is the topic of the next chapter.

5. Modeling acceptability of error in classifiers

5.1 INTRODUCTION

My focus thus far has been on communicating uncertainty: what are the negative consequences when we do not effectively communicate it, and how might we communicate it effectively when we do? However, I believe that effective communication of uncertainty necessitates considering people's preferences from the model on up: that effectively *communicating* uncertainty requires also making decisions that are sensitive to the aspects of uncertainty (or specific types of errors) that people truly care about. In other words, an effective *model*—tuned to users' error preferences—is a necessary part of effective *communication*.

Consider an application where energy appliances are monitored for usage to save on energy costs. Such a system is less useful—frustratingly so—if it consistently confuses two appliances such that the user cannot identify a power-hungry appliance. Patel *et al.* [69] introduced such a system, which uses machine learning to predict the usage of electrical appliances in the home. Their system has an overall accuracy of 85–90% in identifying individual appliances. But how we do know if 85–90% accuracy is acceptable to the users of this application? How much uncertainty is actually tolerable? Also, how sensitive are people to different types of errors—while classifiers in HCI applications are often optimized for overall measures of accuracy like F1 score, people are often differently sensitive to false positives versus false negatives. How can we tell if people prefer higher precision or recall in this space (**T2A**)? Also, would these tolerances change if the same sensing system were used for a different application (e.g., sensing activities of daily living for an aging parent instead of energy monitoring)?

Researchers and developers find themselves trying to extract every bit of inference performance from a system, potentially facing diminishing returns. A follow-on to the Patel *et al.* [69] work by Gupta *et al.* [34] improved the mean classifier accuracy to 94%, but this took several years, significant hardware updates, and identification of new features of interest. Efforts could be made to improve the accuracy even further, but at what point should one focus on the user interface over improving the accuracy of the classifier? Given the increasing prevalence of such systems, we need a systematic way to answer these questions, preferably before even starting to design or improve a classifier for a particular application.

To help researchers address these questions, I developed a model and method for predicting how acceptable users will find the error in systems that use classifiers. My approach takes inspiration from *stated preference modelling* [65] to elicit

This chapter is based on work published at *CHI 2015* with Shwetak Patel and Julie Kientz [50].

This chapter centers on **RQ3:** *Can we tune models to people's error preferences in a simple, lightweight way?*

For reference, **T2** is repeated here:

Using simple survey-based methods that elicit acceptability of error across a space of hypothetical errors in an application, we can **A)** *estimate how much users care about different types of errors in a predictive application (even without ground truth), and* **B)** *express those estimates in the domain language of practitioners building predictive systems such that they can be easily adopted (e.g. as a weighted mean of precision and recall).*

a weighted mean of precision and recall that reflects users' acceptability of error for a system—connecting classifier error to intent to use via acceptability of error and the Technology Acceptance Model (TAM) [98]. My approach focuses on deriving a single weight[1] to describe error preferences in an application, though a full cost matrix for use in cost-sensitive classification [23] could also be derived in a similar manner.

1. For applied machine learning researchers in HCI and UbiComp already using F scores, I believe this approach may be more accessible.

The primary contributions of this chapter center on connecting evaluation of classifiers to acceptability of error, with the aim of predicting the latter on the basis of the former. These contributions are:

1) Formalizing the notion of acceptability of error.[2]

2. In previous work [50] I have called this *acceptability of accuracy* (instead of *error*), using the colloquial sense of *accuracy* in order to reflect users' use of that word. However, I have since adopted the more formally correct *error*.

2) Demonstrating the association between traditional measures of classifier error and acceptability of error (I investigate a class of weighted means of precision and recall that includes the F-measure—commonly used for evaluating novel classifiers in ubicomp, thus satisfying **T2B**—and show how to use acceptability of error to select which of measure to use when evaluating a classifier).

3) Devising and validating a simple survey instrument that developers of inference-based systems can use to help identify acceptable levels of classifier performance before expending the effort to build systems. I assessed the face validity of the survey instrument in four different applications drawn from the ubiquitous computing literature. I also deployed a refined version of the model to demonstrate its predictive validity in the domain of weather forecasting error, showing that we can predict acceptability of error to within one point on a 7-point Likert scale.

This allows us to more systematically answer questions like, *how good does my classifier have to be?* My goal with this work is not to impose further requirements for researchers to demonstrate that their good classifiers are actually good, though I do believe it is possible to make such claims stronger through consideration of acceptability of error. Instead, I aim to provide researchers with the tools to systematically make decisions about how to allocate resources and make design decisions about sensing and predictive systems with user-facing uncertainty: e.g., to think about how to use a seemingly low-performing classifier to build an application with an appropriately fuzzy or broad level of feedback that users will find acceptable, or to refocus resources on the user interface when the classifier is deemed to be "good enough".

5.2 ACCEPTABILITY OF ERROR SURVEY INSTRUMENT

I designed a scenario-based survey instrument to systematically examine the effects of differing classifier accuracies on users' perceptions of those classifiers in the context of specific applications and user interfaces. The basic structure of the survey leads with a description of an application that makes use of a classifier; for example:

For a complete example of an acceptability of error survey, see Appendix A.

> ***Electricity monitor application****: Your residence has been outfitted with an intelligent electricity monitoring system. It is capable of keeping track of how often you use each of your appliances and how much electricity each appliance uses.*

This application description is then followed by a series of accuracy scenarios in which varying levels of performance of the classifier for that system are outlined to participants:

> *Please imagine the following:*
> - ***10 times*** *over a three month period, you used your clothes dryer.*
>   - ***8 of the 10 times*** *that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.*
>   - ***2 of the 10 times*** *that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.*
> - ***2 other time(s)*** *over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.*

This performance scenario lays out several properties of the classifier in bold. In order, they are:

- Real positives (RP); above, the total number of uses of the dryer. This is held constant.
- True positives (TP); above, the number of times the dryer was correctly predicted as having been used.
- False negatives (FN); above, the number of times the dryer was not predicted as being used even though it was.
- False positives (FP); above, the number of times the dryer was predicted as being used even though it was not.

These properties are expressed as frequencies rather than percentages, as work in Bayesian reasoning suggests that people's inferences are better when asked about frequencies rather than proportions [30]. The particular wording for each scenario was developed through pilots on Amazon's Mechanical Turk (http://mturk.com) and in-person.

For a given application, I generate 16 different accuracy scenarios corresponding to 4 levels of recall (0.5, 0.66, 0.833, 1.0) × 4 levels of precision (0.5, 0.66, 0.833, 1.0)[3]. Note that due to the definitions of recall and precision,

$$R = \frac{TP}{TP + FN} = \frac{TP}{RP} \quad \text{(recall)}$$

$$P = \frac{TP}{TP + FP} \quad \text{(precision)}$$

we can calculate all the other values in the above scenarios so long as RP is known (e.g. below RP is fixed at 10).

3. The use of frequencies necessitates some rounding, so some scenarios have only approximately the specified precision/recall.

Participants rate each accuracy scenario on three 7-point Likert-item questions from extremely unlikely to extremely likely:

- I would find the accuracy of this system to be acceptable.
- I would find this system to be useful.
- If available to me now, I would begin using this system sometime in the next 6 months.

These questions correspond to acceptability of error, perceived usefulness, and intent to use (the latter two are adapted from the TAM [19,98])[4].

This structure allows us to generate scenarios for an application with arbitrary accuracy. Essentially, we can sample the space of possible accuracies in an application and then model how this affects acceptability of error. While we have selected particular levels of accuracy here, the scenario-generating code accepts any combinations of levels.

4. Pilot versions of the survey also included ease of use from the TAM, but this question was confusing to users when being asked about a hypothetical system, so I omitted it. Notably, this is consistent with the hypothesis that acceptability of error corresponds to a measure of output quality in the TAM, which there is also not associated with ease of use.

5.3 TESTS OF FACE VALIDITY

I intend this survey to be able to answer several questions about a given application. First, I aim to model acceptability of error based on familiar measures of classifier accuracy. To do that, I derive several measures of accuracy from the precision and recall of each scenario (such as a weighted F-measure) and use these to predict acceptability of error.

Acceptability of error as defined here is intended to correspond to a measure of output quality in TAM2 [98]. Output quality refers to how well a system performs the tasks it is designed for (apart from how useful someone finds those tasks to be in the first place) and has been shown to correlate with perceived usefulness [98]. This leads to the first test of validity:

VT1. Acceptability of error and perceived usefulness should be highly correlated.

Further, per TAM [19,98]:

VT2. Perceived usefulness and intention to use should be highly correlated.

Next, we should not expect two classifiers that have the same quantitative accuracy but which are in different applications to have the same acceptability: users' sensitivity to errors will vary between applications, and the instrument should uncover this; thus:

VT3. The instrument should be sensitive to application: classifiers with similar accuracy for different applications may have different acceptability of error (see **T2A**).

Finally, different types of classification error do not always incur the same cost for users (e.g., the effects of the relative weight of precision versus recall is a well-known problem in information retrieval [79], where it is more important that the top results the user sees are highly relevant than that all relevant results are returned). We should therefore expect the method to be sensitive to such differences in situations where the costs of errors differ. Thus, the fourth test:

VT4. When classifiers with similar accuracy for the same application have different levels of user burden for false positives, the method should be sensitive to this, and reflect it as a different weighting of precision versus recall (see **T2A**).

5.4 STUDY 1: ASSESSING FACE VALIDITY

To validate against the above tests, I used the survey instrument in a survey with four different hypothetical applications inspired by applications found in the ubiquitous computing literature [34,71,103]. This variety was intended to validate **VT3**. The applications include an electricity monitor (introduced above) as well as the following:

***Location tracker**: Your workplace has installed a mobile application on employees' cell phones that can estimate what room at work you or your coworkers are currently in. You can use it to locate a colleague or your supervisor when you are both present at work, for example, to have a quick meeting.*

***Alarm (text message):** Your residence has been outfitted with an intelligent alarm system that is capable of automatically recognizing household members when they enter, without any other interaction. For example, it does not require a password. When a stranger enters the house alone (someone that the system does not recognize), it sends you a text message.*

***Alarm (police):** Your residence has been outfitted with an intelligent alarm system that is capable of automatically recognizing household members when they enter, without any other interaction. For example, it does not require a password. When a stranger enters the house alone (someone that the system does not recognize), it calls the police.*

The two variants of the alarm application are meant to explore two possible extremes of feedback: a relatively low-burden modality (text messages) versus a very high-burden modality (calls to the police). These allow us to validate **VT4**.

5.4.1 Survey structure for data collection

Due to the length of the survey (each application has 16 scenarios, with 3 questions per scenario), I split the survey into two branches (see Figure 5.1). Each participant is randomly assigned to one of two branches: the application branch and the user interface branch, corresponding to **VT3** and **VT4**. Participants in the application branch are asked about the electricity monitor, location tracker, and alarm (text message) applications. Participants in the user interface branch are given the alarm (police) and alarm (text message) applications. Within survey branches, participants were shown each application in a random order. Scenario order within each application was also randomized.

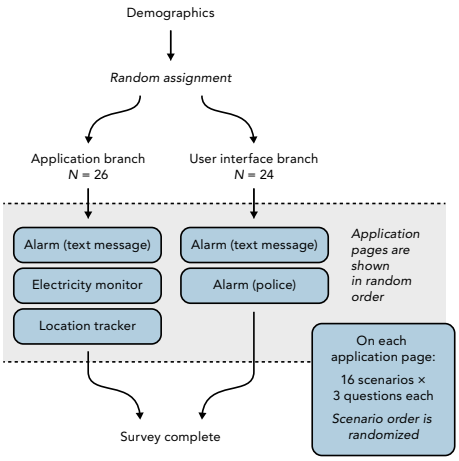


Figure 5.1. Survey 1 structure. Each application page (blue box) corresponds to an instance of the acceptability of error survey instrument applied to a different application.

5.4.2 Participants

Participants were recruited via word-of-mouth, distribution on university mailing lists, and advertisements on Facebook. There were 50 participants, 26 in the application branch and 24 in the user interface branch (this was sufficient to show credible differences in model parameters due to the use of within-subjects design). Participants were entered into a raffle to win one of five Amazon.com gift cards: a $50 card or one of 4 $25 cards. Due to the length of the survey, some participants did not complete the entire survey (11 and 3 in each branch, respectively; each of these participants completed at least one application), which was expected due to its length. I used randomization of scenario order to account for this so that each application still received an adequate number of participants. 50% of participants were female, and 50% of participants in each branch were female.

5.4.3 Model of acceptability of error

To analyze acceptability of error, I posited that acceptability of error may be predicted based on some measure of classifier accuracy. In particular, because precision and recall in these applications are visible to the user, I concentrated on measures based on these (this also aligns with the common use of F-scores in ubicomp literature). I did not consider measures that involve true negatives, as it is not clear in the scenarios that true negatives are meaningful to users. For example, what does it mean to a user that their alarm correctly did not go off whenever no one was breaking into their home? Rather, the user cares about precision: when I actually see an alarm, how likely is it genuine? This also offers a simpler starting point to model.

Thus, I first consider the weighted F-measure, which is equivalent to a weighted harmonic mean of precision and recall. Note that it can be considered a part of a larger class of weighted power means of precision and recall:

$$M_{p,\alpha}(P,R) = [\alpha P^p + (1-\alpha)R^p]^{\frac{1}{p}}$$

In this class, p specifies the type of mean; for example:

$$\begin{aligned} M_{-1,\alpha}(P,R) &= \left(\frac{\alpha}{P} + \frac{1-\alpha}{R}\right)^{-1} && \text{(harmonic mean)} \\ M_{0,\alpha}(P,R) &= P^{\alpha}R^{1-\alpha} && \text{(geometric mean)} \\ M_{1,\alpha}(P,R) &= \alpha P + (1-\alpha)R && \text{(arithmetic mean)} \end{aligned}$$

The parameter $\alpha \in [0,1]$ specifies a relative weighting of recall and precision; when $\alpha = 0.5$, both are weighted equally; when $\alpha < 0.5$, recall is weighted higher than precision; and when $\alpha > 0.5$, precision is weighted higher than recall. In this class, $M_{-1,\alpha}$ is equal to $1 - E_{\alpha}$ (van Rijsbergen's Effectiveness measure [79]), or the F_{β}-measure

where $\alpha = 1/(1+\beta^2)$; thus $M_{-1,0.5}$ is the familiar F1 score. $M_{0,\alpha}$, the geometric mean, is also known as the G-measure [75].

I consider this larger class of measures so that there is a systematic way to ask both whether harmonic mean (i.e., F measure) corresponds most closely to how people judge acceptability of error for these applications (by determining p) and so that it is possible to estimate whether for a given application, people value precision or recall more highly (by determining α for that application).

I conducted a mixed-effects Bayesian logistic regression of acceptability of error against three different weighted power means of precision and recall (harmonic, geometric, and arithmetic). The model was as follows:

$$\begin{aligned} \text{logit}(\mu_{i,j,k}) &= \beta_{0,i} + \beta_{1,i} M_{p,\alpha_i}(P_{i,j}, R_{i,j}) + U_k \\ acceptability_{i,j,k} &\sim \text{Bernoulli}(\mu_{i,j,k}) \end{aligned}$$

For respondent k on scenario j in application i, with p drawn from a categorical distribution over (–1,0,1) corresponding to the aforementioned three types of means. Here, $acceptability_{i,j,k} = 1$ when a participant rates the acceptability of error for that scenario as *Slightly likely* or higher and $acceptability_{i,j,k} = 0$ otherwise.[5] U_k is the random effect for participant k. I used the following uninformed priors:

$$\begin{aligned} \beta_{0,i}, \beta_{1,i} &\sim \text{Normal}(0, 1E12) \\ \alpha_i &\sim \text{Uniform}(0,1) \\ U_k &\sim \text{Normal}(0, 1/\tau) \\ \tau &\sim \text{Gamma}(0.001, 0.001) \\ p+2 &\sim \text{Categorical}(1/3, 1/3, 1/3) \end{aligned}$$

This model allows us to separately estimate α_i for each application i. In addition, the posterior distribution of p will give us an estimate for how believable each type of mean is as a predictor for acceptability.

I take a Bayesian approach rather than a null-hypothesis significance testing (NHST)-based approach in modeling acceptability for several reasons. First, it yields a richer estimation of the parameters of interest. In particular, it allows us to estimate a complete posterior probability distribution of α for each application, rather than just a (point) maximum likelihood estimate. Second, as part of my goal is to propose methods of classifier evaluation that others can build upon, a Bayesian approach is a natural fit: posterior distributions of parameters from the model (and hopefully in the future, others') can be used to inform prior distributions in future work.

I adopt Kruschke's [56] approach to Bayesian experimental statistics. In particular, I examine 95% highest-density intervals (HDIs) of posterior distributions to estimate

5. While I considered using an ordinal or a multinomial logistic regression instead of a binomial regression, ultimately the question when evaluating a classifier here becomes "how many people said the accuracy was acceptable at all?", in which case this threshold would be applied after regression anyway, so the simpler model suffices while invoking fewer assumptions.

Note that section 5.5.1 uses an ordinal regression model to better take advantage of a more sparse sampling of the precision/recall space.

credible differences between parameters (as opposed to an NHST approach of a 0.05 p-value threshold on the distribution of a test statistic).[6]

6. Where possible I also ran similar more traditional NHST models and saw similar effects.

5.4.4 Results

The posterior distribution of p represents an estimation of which measure best approximates acceptability of error. For these applications, $p = 0$ (geometric mean) is most credible ($P(p = 0) = .81$), suggesting the G-measure may more closely correspond to users' estimations of accuracy here. It is worth nothing that F-measure was more probable than arithmetic mean, and had only moderately less believability than G-measure ($P(p = -1) = .17$, Bayes Factor = 4.7). Figure 5.2 plots the proportion of people who rated the acceptability of error at each level against the weighted geometric mean for that application derived from the model. Higher weighted mean is clearly associated with greater acceptability of error. The rest of the results are broken down based on the validity tests outlined above.

VT3. The instrument should be sensitive to application: Classifiers with similar accuracy for different applications may have different acceptability of error.

Confirmed. See Figure 5.3: In the application branch of the survey, α for the electricity and location applications were both ~0.5, but for alarm (text message), α was ~0.3. The differences between alarm (text message) and electricity monitor ($\alpha_i - \alpha_j = -0.19$, 95% HDI: [−0.30,−0.09]) and between alarm (text message) and location tracker ($\alpha_i - \alpha_j = -0.19$, 95% HDI: [−0.32,−0.07]) were credible on a 95% HDI, suggesting that the tool can be sensitive to differences in preferences between applications. In addition, β_1 for all applications on both branches was credibly different from 0 its 95% HDIs. While it varied by application, $\beta_1/100$ typically corresponded to an odds ratio of ~1.2, or a 20% increase in the odds that a person finds the accuracy of a system acceptable for every 0.01-point increase in G-measure.

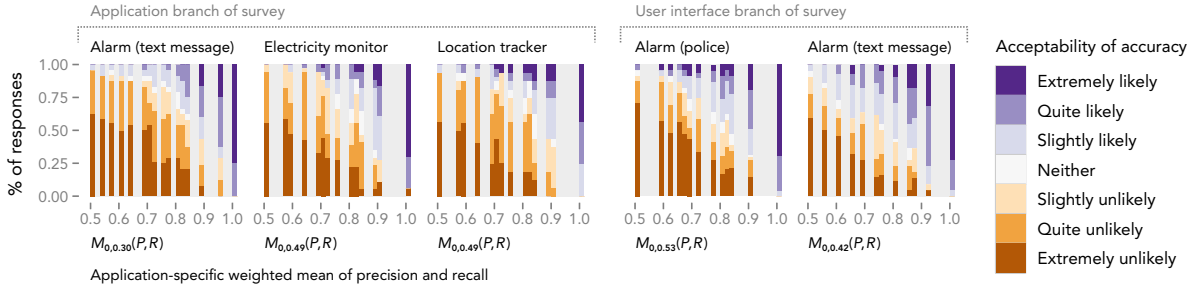


Figure 5.2. Acceptability of error from the survey results plotted against each application's weighted geometric mean of precision and recall from the model. Optimizing this mean has the effect of also optimizing an application's acceptability of error.

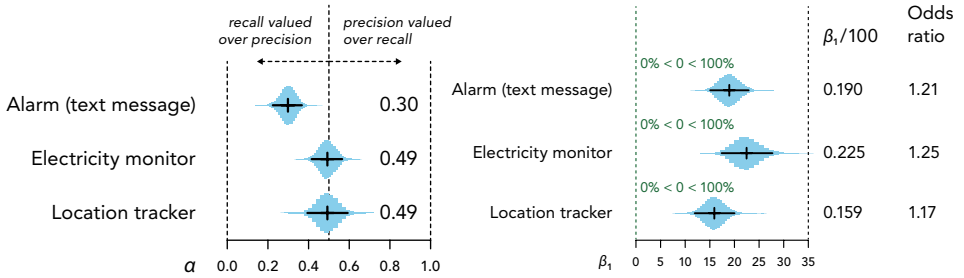


Figure 5.3. Posterior distributions of α and β1 for the application branch of the survey. Mean and 95% HDI are indicated. Note the sensitivity of the model to different preferences of precision versus recall between applications.

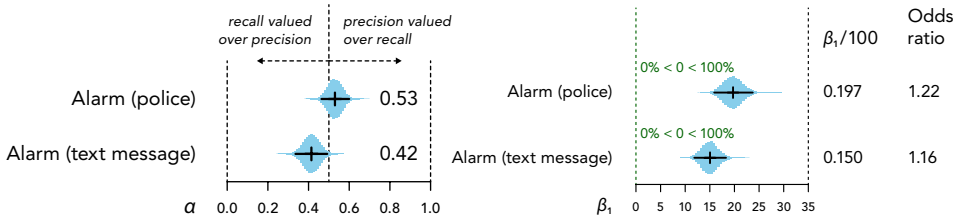


Figure 5.4. Posterior distributions of α and β1 for the user interface branch of the survey. Mean and 95% HDI are indicated. Note the sensitivity of the model to different preferences of precision versus recall between feedback types.

VT4. When classifiers with similar accuracy for the same application have different levels of user burden for false positives, the method should be sensitive to this, and reflect it as a different weighting of precision versus recall.

Confirmed. See Figure 5.4: In the user interface branch of the survey, the alarm scenario had $\alpha = 0.53$ for police compared to the much lower $\alpha = 0.41$ for alarm with text message, and these differences were credible on 95% HDI ($\alpha_i - \alpha_j = -0.12$, 95% HDI: [−0.22,−0.01]). This demonstrates that the relative weighting of recall and precision for the same classifier on the same application—but with different feedback—can be quite different. Here, a more lightweight form of feedback (text messages) leads users to value recall over precision—that is, they are much more willing to tolerate false positives in order to obtain a higher true positive rate.

Note also that the bulk of the posterior distribution of α (81%) for alarm (police) is also greater than 0.5 (although 0.5 is not outside its 95% HDI), giving us some evidence that participants here valued precision over recall. This is as we would expect given the type of feedback: a false positive is costly if it results in a call to the police.

VT1. Acceptability of error and perceived usefulness should be highly correlated.

These measures were highly correlated according to the Spearman rank correlation coefficient ($\rho = 0.89$, $p < 0.001$), suggesting the validity of the inclusion of acceptability of error as a measure of output quality in the TAM.

VT2. Perceived usefulness and intention to use should be highly correlated.

These measures were also highly correlated ($\rho = 0.85$, $p < 0.001$).

5.5 STUDY 2: ASSESSING PREDICTIVE VALIDITY

To assess the predictive validity of the tool, I conducted a survey of people's perceptions of weather forecasting apps and websites. Weather prediction is a system where people are regularly exposed to the effects of prediction accuracy (e.g. failing to bring an umbrella when it rains) without knowing the precise accuracy of the system, as we might expect in other user-facing classification systems, making it a good candidate for validation. I obtained ground truth data of precipitation predictions from various weather forecasters in a major metropolitan area of the United States over the time period from Sept 1, 2013 to Aug 31, 2014[7]. I focused on one city, as people in different climates may have different preferences for precipitation accuracy. This survey had two parts:

7. This data set was obtained from http://forecastwatch.com, an independent company that tracks the accuracy of weather forecasters.

Part 1: Existing acceptability (ground truth). I asked participants to specify which weather forecasting apps and websites they currently use and to rate each on acceptability of error of precipitation prediction over the last 30 days and the last year. I also included usefulness, ease of use, and frequency of use questions for validating against the TAM. The part of the survey again used 7-point Likert items, but used the anchors *strongly disagree*/*strongly agree* instead of *extremely unlikely*/*extremely likely*, as these statements were not predicting hypothetical use but describing existing opinions. Unlike with the survey tool, these questions do not specify the accuracy of the systems in question to participants. However, since I have the ground truth of the predictive accuracy of these systems over both time periods in question, it is possible to model these responses in a similar manner to the hypothetical accuracy survey without the caveat that people are responding to hypothetical levels of accuracy.

Part 2: Hypothetical acceptability (the survey tool). The survey randomly selected one application or website from Part 1 that the participant currently uses and generated a variant of the survey instrument for that application or website. Participants were asked to imagine the randomly selected weather app had the accuracy described in each scenario (thus making the scenario more concrete), then to rate acceptability of error, usefulness, and intent to use. Each scenario began, "15 days in a 30-day period, it rained" (i.e., real positives were fixed at 15). As before, the scenario specified TP, FN, and FP (and additionally TN, as there was a fixed time interval and prevalence) using four statements like "13 of the 15 days that it rained, the weather forecast had (correctly) predicted that it would rain that day." I used the same precision and recall levels as before, but instead of giving all 16 scenarios to each participant, each participant saw 8 randomly selected scenarios to reduce survey length.

5.5.1 Revised model of acceptability of error

Due to the reduced number of scenarios shown to each participant (potentially sampling over a smaller space of possible answers on the Likert scale for any individual participant), I used an ordinal regression instead of a binomial regression.[8] This revised model assumes the same latent variable representing acceptability is associated with the ordinal ratings of acceptability of error in Part 1 and in Part 2.

8. In contrast to the simpler model of section 5.4.3.

Besides the use of ordinal instead of binomial regression, this model also includes two additional parameters. I added a fixed effect of survey part (*ground truth last 30 days*, *ground truth last year*, or *hypothetical*), $\beta_{2,i}$, to estimate whether people systematically over- or under-estimate their actual acceptability, as we might expect if (for example) people answering "extremely likely" are likely to give a lower rating of acceptability of error when answering about a system they have experienced. I also added a scaling parameter, ζ_i, (also varied by survey part) to estimate whether people are more or less sensitive to changes in accuracy in real systems versus the hypothetical scenarios. By modeling these effects, we can use an individual's predictions about their own acceptability in hypothetical scenarios to estimate how acceptable they will actually find the accuracy of those systems to be if they used them. The model specification is:

$$\text{logit}\big(P(Y_{i,j,k} \leq l)\big) = \frac{\theta_l - \beta_1 M_{p,\alpha}\big(P_{i,j}, R_{i,j}\big) - \beta_{2,i} - U_k}{e^{\zeta_i}}$$

For acceptability level l for respondent k on scenario j (or forecaster, in Part 1) in survey part i. Scale and location parameters for the hypothetical part of the survey are fixed (where $i = 3$), $\beta_{2,3} = 0$ and $\zeta_3 = 0$, so that the parameters from the ground truth parts of the survey ($\beta_{2,1}$, $\beta_{2,2}$, ζ_1, and ζ_2) can be interpreted as shifts in the location and scale of a person's rating when they go from hypothetical scenarios to actual experience.

I use leave-one-participant-out cross-validation to assess the predictive validity of the model. In each fold, I fit a model with all participants' responses on Part 2 (hypothetical acceptability), but use all participants' responses except one to estimate the bias of hypothetical responses versus acceptability of known systems (Part 1). I then use the responses of the left-out participant in Part 2 to predict how they would rate the randomly-selected weather forecaster they saw in Part 2 on Part 1 of the survey based on the known accuracy of that forecasting app.

This mimics the situation where the accuracy and acceptability of an existing

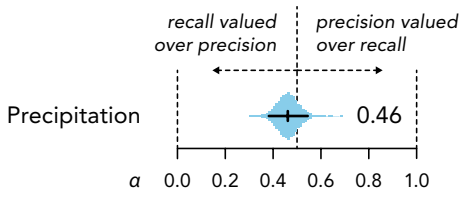


Figure 5.5. Posterior distribution of α for precipitation prediction with mean and 95% HDI. There is some evidence for wet bias: 82% of the distribution lies below 0.5.

system is known, but the real-world acceptability of future (possibly better) systems is unknown. This scenario might arise (for example) if researchers in an area wish to set future targets for accuracy; such a model would provide the ability to predict acceptability levels in a population based on a combination of people's opinions of existing systems and their ratings of hypothetical systems.

5.5.2 Participants

Participants were recruited and compensated as in Study 1. There were 22 participants, 55% of which were female.

5.5.3 Results

The model had a cross-validated mean absolute error (MAE)—the mean difference between the predicted acceptability and actual acceptability in points on the Likert scale—of 0.93, suggesting the predictions are generally within one point of the actual value on the scale. This is promising, though it is worth noting that people's ratings of actual systems were generally tightly clustered in the "acceptable" range (the MAE of using the median category as a predictor was 1.09).

This tight clustering was also reflected by the model. While weighted precision/recall had a credible effect ($\beta_1 = 13.4$, 95% HDI: [9.92, 16.7]), scale parameters for the ground truth data indicated that ground truth responses were credibly less variable ($\zeta_1 = -0.437$, 95% HDI: [−0.829, −0.066]; $\zeta_2 = -0.346$, 95% HDI: [−0.673, −0.003]). These coefficients suggest that responses in the ground truth data had about 60% the variance of the hypothetical data. In other words, people were less sensitive to changes in accuracy in the systems they used than they predicted they would be in the hypothetical survey. People also tended to underestimate the acceptability of precipitation predictions in hypothetical scenarios, with ground truth responses having credibly higher ratings for the same accuracy ($\beta_{2,1} = 1.59$, 95% HDI: [1.1, 2.13]; $\beta_{2,2} = 3.82$, 95% HDI: [2.94, 4.8]).

The model found some evidence of wet bias [85] in participants' preferences: the estimated α was 0.461 (95% HDI: [0.382, 0.545]) with 82.3% of the distribution lying below 0.5 (see Figure 5.5). This leads some credence to the idea that people may weight recall higher here—desiring forecasters to catch more instances of rain in the forecast at the expense of making more false predictions of rain. I expect the prevalence of this bias to vary by climate, so make no claims to generalizability beyond the city tested in. I also asked people to state whether they thought it was worse when the forecast "does not call for rain, but it does rain" (FN) or when "calls for rain, but it doesn't rain" (FP), and 88% considered the former worse, consistent with a higher weight on recall, further validating the model.

As before, acceptability in the hypothetical survey was highly correlated with usefulness ($\rho = 0.98$, $p < 0.001$) and usefulness with intent to use ($\rho = 0.95$, $p < 0.001$). There were also significant correlations between acceptability of error and usefulness in the ground truth survey, though less strong ($\rho = 0.37$, $p < 0.001$), which is to be expected in real-world systems (where other concerns such as usability and convenience have additional salience over accuracy). Importantly, there were no significant correlations between acceptability of error and ease of use in the ground truth survey ($\rho = 0.13$, $p = 0.13$), but there were moderate correlations between ease of use and usefulness ($\rho = 0.55$, $p < 0.001$)—as predicted by the TAM—suggesting that acceptability of error is a separate construct from ease of use and is more related to output quality and usefulness in the TAM, as hypothesized.

5.6 DISCUSSION AND IMPLICATIONS

In this section I discuss implications of the survey instrument for estimating acceptability of error and its potential uses. This research has broad implications, from deciding how to evaluate classifiers in user-facing systems, to selecting user interfaces and feedback for a new system, to allocating research resources.

5.6.1 What can we say absent a user interface? Selecting an objective function

Given a new classifier, typically, we might tune this classifier to optimize the F_β-measure (where β is usually 1). However, even without acceptability of error ground truth, this instrument can be used to decide a more appropriate objective function to optimize during learning (e.g. an F or G measure with a particular weight). While the actual acceptability will not be known (because we cannot estimate shifts in location or scale of real-world acceptability without data from actual use), I believe that this estimated objective function will correlate with real-world acceptability of error more closely than (say) F1 score (**T2A**). More broadly, I believe researchers should consider whether F1-measure truly matches their evaluation goals before employing it on user-facing systems.

5.6.2 Selecting a user interface to build: The potential of a low-performing classifier

As researchers in HCI and ubicomp, we often find ourselves asking, is this classifier good enough for our users? Indeed, I can recall several conversations with colleagues working on classifiers for various problems wherein someone asserted that the classifier was not good enough—and yet, the system had no user interface to speak of. If we have a classifier with (e.g.) better precision than recall, we can use the acceptability of error instrument to test out several hypothetical user interfaces or applications for a given classifier, and then build the application in which

people weight precision as more important than recall (or vice versa, as called for by the results from the instrument). This provides a way to increase the chances of building an acceptably accurate user-facing system given a classifier with known shortcomings. Given the potential for lower-burden, fuzzier feedback to improve the acceptability of error of a system, it may be premature to rule out a weak—but adequately-performing—classifier without investigating acceptability of error for potential instantiations of its user interface.

A lower performing but still acceptable classifier might also be used to preserve privacy or plausible deniability, which I believe this approach can help uncover. More simply, the lower performance classifier might be the easiest and cheapest to implement given system's computational capabilities. Knowing how accuracy trades off against acceptability would enable researchers to make these types of judgments more systematically.

5.6.3 Same sensor, different application: performance may not transfer

In a similar vein, a classifier that appears quite accurate for its domain may not have acceptable accuracy depending on what kind of application it is built into. For example, one might consider building many different types of systems on top of infrastructure-mediated sensing (e.g. sensors that can disaggregate energy [34] or water [27] use by appliance). The obvious example is an application for identifying high-cost appliances on a utility bill. However, a parent might also wish to use such a system to track TV usage of their child. While a certain level of false positives in tracking energy use of appliances seems unlikely to cause large discrepancies in finding high-cost devices, a few false positives in TV use may be the source of arguments between parents and children about TV-time quotas. We could more systematically investigate these intuitions by fitting a model of acceptability of error to each of these applications. This would allow us to decide whether our classifier is adequate for each application use case.

5.6.4 Predicting future acceptability and setting targets

Given actual use of a classifier with known accuracy, acceptability ratings of that accuracy (collected as I did with the weather data), and results from the survey instrument, we can estimate the relationship between hypothetical and actual acceptability, as demonstrated in the case of weather prediction accuracy. In this case, we can actually use hypothetical ratings to estimate the acceptability of error for potential future classifiers that are more accurate than existing ones, and use this model to set targets for desired classifier accuracy or to identify when we have reached a point of diminishing returns.

5.6.5 Training a new model: predicting when to predict

Many systems require an initial training period with a new user before they can make predictions (e.g., the Nest thermostat, Belkin Echo electricity/water monitoring); such systems wait until they have built a good personalized model. But how long should this training period be? The first impressions made by poor predictions are likely to sour users on a system. Given a model of acceptability of error for an application, one could set a desired threshold of acceptability (e.g., as a proportion of the user base), and use this to determine at what point a system should switch from training to prediction.

5.6.6 Expanding to other application domains

Thus far I have examined four specific application domains: electrical appliance detection, person location within an office, home alarms, and precipitation prediction. I chose applications that would be broadly applicable to a general audience (simplifying recruitment) and that could be used to validate the instrument. However, there are many other domains that can still be explored. For example, health sensing and recognition of daily activities for older adults are two popular application areas within HCI and Ubicomp. These types of applications are often only useful to certain subsets of people (e.g., someone with a specific health condition or someone caring for an older person), and thus if these domains are tested, the surveys should be targeted toward these specific populations rather than a general population (a primary reason I did not test them here). I suspect that health and older adult applications might require a higher level of accuracy, but that the user interface will again matter greatly. This is something the survey tool is designed to determine.

To facilitate adoption, I plan to open source code for generating the survey based on desired precision and recall levels and code for fitting the model. I envision building an online repository of application examples and resulting data that can be used as guidelines to others wanting to build classifiers in a given space. For example, if someone is interested in exploring a new sleep sensor, they might look up data for similar applications in that domain and find that they need to aim for about 90% accuracy (as measured by some particular measure of accuracy, like weighted G-measure). This could also serve as a sort of "grand challenges" list for the community to help people building classifiers find interesting problems worth solving, rather than spending resources on areas with diminishing returns. At some point, resources on any given application may be better spent on improving the user interface or on another domain altogether.

5.6.7 Recommendations on applying the survey to research

My experience in conducting this research leads to several recommendations for researchers hoping to apply a similar approach to their own applications and classifiers. I recommend presenting each user with at most 8 accuracy scenarios (as in the weather application), as I received feedback that the original survey (with 16 scenarios) was a bit long. I also recommend including at most two applications at a time, as the survey with three different applications had a higher rate of partial completions (11/26 compared to 3/24 in the two-application branch). Note that due to its design, a small number of participants (here, ~20–25 per application) is sufficient to achieve credible estimates of the model parameters from the survey tool.

In addition, although the current tool uses written scenarios, researchers should consider other forms of representation of the system, such as visual screen mockups, storyboards, or video scenarios to help explain the intended use. Deployment on Mechanical Turk offers another approach, where each scenario can be made a single, small task, mitigating fatigue.

5.6.8 Limitations and future work

I believe that my proposed approach to assessing acceptability of error gives researchers an easy-to-use method for assessing acceptable error levels for a given classifier and interface. However, there are some limitations. First, the models are typically application-specific, and expansion and testing in more domains is necessary. A good next step to address this limitation would be to test on more systems: for example, to simulate varying accuracies within a home electricity monitoring system and see whether people's perceptions of acceptability of error can be predicted using the acceptance of accuracy survey (similar to how I validated the precipitation prediction model). I also believe that model estimates from previous, similar applications can inform future models (and here, the Bayesian approach used can facilitate this). Finally, as an initial test case for this approach the survey thus far is geared toward evaluating the effect of precision and recall in binary classifiers. Further work is necessary to see how (e.g.) true negatives affect perceptions, to incorporate a broader set of classifier evaluation measures (c.f. [72]), or expand to contexts beyond binary classification.

5.7 CONCLUSION

I believe that communicating uncertainty effectively requires working across the full stack of a predictive system, from the model all the way up to visualization. A model that is not in tune with users' preferences and expectations cannot be saved by slapping a more effective presentation of uncertainty on top. In this chapter, I proposed one method for tuning binary classifiers to users' preferences for error,

inspired by the Technology Acceptance Model (TAM). The goal of this work was to connect work in classifier evaluation to notions of technology acceptance, which I did through the introduction of a construct I call *acceptability of error* and a survey instrument I constructed and validated. I believe I have demonstrated a lightweight way to evaluate and tune models to be sensitive to errors people actually care about, and thereby to make it easier for applied machine learning systems to be built that make decisions more in tune with users' preference for error. This, alongside effective communication of uncertainty, is an important part of designing systems with user-facing uncertainty—complementing the previous two chapters. With the addition of this piece, I aim to lay the foundation for tackling communication of uncertainty across all levels of a predictive system.

6. Discussion and future work

6.1 THE VIEW FROM THE START

6.1.1 Considering RQ1: How does the lack of communication of uncertainty affect trust in a simple, real-world sensing system (body weight scales)?

Building on the survey-based work of Lim and Dey [60], I have found evidence for issues of trust and abandonment precipitated by missing uncertainty information in a real-world system: the body weight scale. Statistical misconceptions, like confusing bias with variance or being unable to discount short-term variation, are only exacerbated by systems that provide estimates one point at a time. Failing to communicate uncertainty has *real* consequences on trust, and can lead people to abandon their scales as being "too inaccurate". I have found that these are not just theoretical problems, and as we expand the reach of estimation and prediction in our daily lives, we must do better to communicate uncertainty effectively to end-users.

6.1.2 Considering RQ2: Can we build visualizations of uncertainty in continuous predictions (bus arrival times) that people do understand?

I believe the answer to this question is yes. I uncovered unaddressed needs in realtime transit prediction that can be served through communicating uncertainty, and then through surveys and iterative design I developed a redesigned interface for communicating probabilistic predictions of realtime transit arrival times. Building on the success of discrete visualizations of uncertainty [29,30,37], this interface employed a novel discrete visualization of uncertainty for a continuous prediction I call *quantile dotplots*. Tested on over 900 existing users of a realtime transit prediction app, I found that the precision of people's probabilistic estimates improved by around 15% using quantile dotplots, and their confidence in their estimates also improved. This suggests that this interface and visualization technique *can* help people understand probabilistic predictions.

6.1.3 Considering RQ3: Can we tune models to people's error preferences in a simple, lightweight way?

Building on the idea of stated preference modeling, my acceptability of error survey instrument demonstrates one way to associate to traditional measures of classifier error to technology acceptance. My approach facilitates model tuning and evaluation in a simple, lightweight way by yielding a single weighted mean of precision and recall that reflects users' error preferences. I also provide initial validation for this survey instrument. I believe that the basic structure of this approach—connecting classifier evaluation to intent to use through acceptability of error—can be

easily applied to other classifier evaluation metrics, or even to estimating full cost matrices in cost-sensitive classification. The approach itself involves a randomized survey instrument that is straightforward to tailor to new applications. This approach would allow effective communication of uncertainty to be considered across the entire prediction stack, from modeling to output.

6.2 LESSONS LEARNED

6.2.1 The importance of discrete / frequency-based representations; or, 10 out of 10 doctors recommend saying "10 out of 10" instead of "100%"

The quantile dotplots I developed for realtime transit prediction represent another technique for (and some more evidence indicating the value of) representing uncertainty using discrete outcomes. I employed a similar frequency-based approach to communicating uncertainty in the scenarios used to elicit acceptability of error, to some success. Across many domains [29,30,37], this approach is slowly becoming recognized for its ability to communicate uncertainty to lay people in a way they can better understand. My work extends this success both by demonstrating its value in new domains, and by introducing a novel discrete visualization of continuous probability distributions (quantile dotplots) that can improve the precision of people's probabilistic estimates.

6.2.2 The importance of a common language to communicating uncertainty; or, our precise language must have low variance to be consistently understood

Vocabulary—and the difficulty of nailing down words for the just the core concepts of error, like *bias* (accuracy, consistency, ...) and *variance* (precision, consistency (but not that kind of consistency!), ...), let alone more complex concepts like inter-scale reliability—underscores many of the issues of communicating uncertainty effectively. Without even the knowledge of the *concept* of precision as distinct from a broad, colloquial notion of "accuracy", how can we hope to establish a common language to *talk* to our users in (however we define *talk*)? Over the course of my work I have increasingly become convinced that natural language is some piece of that system-user conversation about uncertainty, and also that a more systematic understanding of the lay language of uncertainty is necessary to make that conversation flow.

I have uncovered glimpses of what those conversations should look like: the particular vocabulary recommendations derived from my survey of weight scale users, and some of my proposals for the use of natural language communication of uncertainty (in weight scales, and also in some of my prior work in sleep sensing not discussed herein [46]). However, there is a lot of work left to get this right.

Distinguishing bias from variance is the easiest piece to this problem, and it is already hard. How should we—or should we—distinguish epistemic uncertainty (e.g., uncertainty in the location of a parameter) from aleatory uncertainty (e.g., uncertainty resulting from another draw from a distribution, as in a prediction)? How do these interact with discrete representations of uncertainty—does a discrete representation of epistemic uncertainty even make sense to people?

One way forward might be to squash all of that uncertainty together and simply try to make predictions. Parameter uncertainty (say a posterior distribution for the mean arrival time of a bus) can be sampled over and combined with a residual distribution (the variance, skew, etc. of actual outcomes around that mean—the aleatory uncertainty) to generate predictive distributions that reflect both epistemic uncertainty and aleatory uncertainty.[1] I suspect—but have no direct evidence for as yet—that focusing on predictions is one way to sidestep the need for users to understand the difference between aleatory and epistemic uncertainty (or even more complex components of a model), so long as uncertainty is included in the predictions (so as not to erode trust when a prediction goes awry). This is something of a recasting of the argument I made that an improved model for realtime transit prediction might reduce the need to communicate factors like the bus delay or the time since last update. A prediction is a *concrete* thing, with uncertainty that can be communicated as discrete outcomes in a way that makes sense viscerally ("these twenty possible busses..."). Why try to explain the meaning of a confidence interval when we can just make a prediction that is relevant to the user's actual decision (and which itself accounts for the uncertainty of that confidence interval, and all other uncertainty in the model besides)?

1. This is a common way to make predictions from Bayesian models, called a *posterior predictive distribution*; see, for example, Chapter 3 of McElreath [64].

Finally, reflecting on the language of uncertainty forces me to reflect on the difficulty of even discussing research in this space with others. Apart from any studies I have run, the lack of a more accessible language for communicating uncertainty makes it a difficult topic to communicate *about*. And yet, communicating uncertainty—and communicating about uncertainty—is increasingly important to our everyday lives.

6.2.3 The importance of reflecting uncertainty as a core value in sensing and prediction; or, ignorance is not bliss

Communicating uncertainty is invading our everyday lives because of the increasing prevalence of applications built on sensing and prediction. Across all of my work, I have found ignorance not to be bliss: the point estimate without uncertainty conveys false precision nearly wherever it rears its head (in weight scales, in transit prediction, in weather prediction, and in innumerable other domains I have discussed with domain experts—particularly in subdomains of health). Present in a

system, that false precision erodes trust. To push back against the primacy of the "best guess", communicating uncertainty needs to become a core value in sensing and prediction—it should come first and by default, and *hiding* it should be an intentional design decision, not the other way around.

6.2.4 The importance of considering model, UI, and representation together from the start; or, it's certainly uncertainty that shouldn't be added retroactively

Making uncertainty the default presents not only questions of how to communicate that uncertainty, but also how the model should deal with that uncertainty. While my work on acceptability of error underscored the need to consider what errors people care about when building a model, this is not the only way that the user interface can dictate the type of model needed. Certain types of models cannot even produce predictive distributions, only point estimates (and unfortunately in many cases where models could easily give a predictive distribution, modelling software may not make such distribution easy to generate by default). Thus, a conscientious designer has technical constraints of software added on top of user-driven constraints like errors people care about or how well they understand different representations of uncertainty. All of these must be considered from the start of the design process in order not to end up with a model that defies users' expectations of error, or produces estimates that could not be communicated effectively to them even if it matched their acceptability of error.

6.3 META-LESSONS LEARNED

6.3.1 Behavioral economics has something to say about HCI

Somewhere in all of this work is a lesson in the value of behavioural economics to the field of human–computer interaction, by counterexample. My work in acceptability of accuracy, and in evaluating understandability of predictive distributions, could be recast as decision problems and conducted as behavioral economics experiments (e.g., simulations in which people make decisions with payoffs, and are able to learn from their decisions to attempt to improve over time). This is an approach that, in retrospect, I wish I had known more about five years ago. I think that such an approach can push lab studies closer to the real world, as another point on the spectrum leading up to a deployment study. I have already begun a follow-up project to the transit arrival prediction work taking such an approach, to see if different visualizations of uncertainty can improve people's decisions in bus-catching simulations, with costs and rewards for getting coffee, missing the bus, or making a meeting on time. Many of the most interesting questions in HCI come down to how people make decisions with computing systems, and I think

HCI should continue its tradition of methods appropriation to tackle these sorts of questions.

6.3.2 Bayesian statistics has something to say about HCI *and* about user-facing uncertainty

About halfway through my PhD—possibly jarring to a reader of this thesis—I began adopting Bayesian statistical methods in my work. These methods vastly expanded the statistical language I could use to tackle research questions. It made casting the problem of acceptability of error as a weighted mean of different types of errors straightforward. It made decomposing error in people's estimates from predictive distributions of bus arrival time into bias and variance straightforward. It made understanding statistical problems as problems of estimation—rather than hypothesis testing—straightforward.

However, it was not until late in my degree that I connected my attraction to the richness of this statistical language back to my fascination with communicating uncertainty. Bayesian statistics gives a consistent way for uncertainty to cascade through statistical models, but it also gives an interpretation to that uncertainty that I think people have some chance of grasping: *given what I know, here is the probability of X*. This straightforward understanding can supplant the strange double-negative logic of frequentist constructs like *p* values and confidence intervals. In trying to find ways to communicate statistical results to lay people, I realized that even scientists struggle with frequentist statistics. Perhaps we are all lay people with respect to visceral understanding of probability, and perhaps Bayesian statistical approaches can help all of us.[2]

2. Hence my argument for Bayesian statistics as *researcher-centered* statistics [49].

6.4 FUTURE WORK

I have outlined some future work below not alluded to in the lessons learned above.

6.4.1 In communicating uncertainty in weight scales

As described at the end of Chapter 3, the immediate future work to be completed in communicating uncertainty in weight scales consists of testing my proposed weight scale redesign in a deployment study. While I believe the work I have done has set a solid foundation of design principles (realized by the design I have proposed), there remains work to be done in controlled evaluation of those principles. I have built an implemented a system to carry out this evaluation, and plan to do so.

6.4.2 In visualizing uncertainty as discrete outcomes (e.g., quantile dotplots)

While my work has found that quantile dotplots improve people's estimates of uncertainty in the transit arrival prediction domain, I am also pursuing a simulation

study to evaluate how well people make decisions given different visualizations of transit prediction uncertainty. I also plan a deployment study based on a modified version of the existing OneBusAway app to test visualizations of uncertainty, like quantile dotplots, in this space. Beyond that domain, there are many other domains to explore that require communicating uncertainty in continuous predictions to end users that may be amenable to techniques like quantile dotplots. There are also interesting questions about how to communicate uncertainty in more complex types of predictions; for example, a joint probability distribution of predicting arrival time of one bus and the departure of another could be used to help people make a transfer, but may involve a more complex visualization than a univariate predictive distribution.

6.4.3 Further predictive validation In acceptability of error

I believe that the next step in my acceptability of error work is to pursue further predictive validation in a manner that admits better control over the types of errors people experience (in contrast to the weather prediction context, which relied on weather forecasters making errors—which created problems when forecasters did not make as substantial errors as we might need to estimate differences in acceptability).

Disaggregated energy monitoring is an ongoing area of research in ubiquitous computing [34,69], where the goal is to monitor energy use of different appliances in the home with minimal hardware installation (e.g. by installing a system at a single point on the power line rather than at every outlet). Naturally, such an approach is not perfectly accurate. I believe that a domain like could be used to conduct further predictive validation of my acceptability of error instrument. This offers a domain closer to the original motivation of the instrument than weather forecasting, as well as an opportunity to conduct a real-world deployment that is closer to what might be necessary as an example for others wishing to adopt these techniques in practice. In such a domain, a deployment study could be run using a ground truth sensor (e.g., a meter installed directly on the outlets for various appliances), and then error injected into the output given to users. Then, one could compare the predicted acceptability of error from an initial survey to participants' actual acceptability of error when they experience errors during use.

7. Conclusion

In my thesis I have integrated two parallel approaches to improving sensing and predictive systems through consideration of uncertainty: first, by asking whether and how we should communicate uncertainty to end users, and second, by developing a survey instrument to help inform the design of both algorithms and user interfaces for a predictive system. My work was driven by three questions: do we need to improve communication of uncertainty (**RQ1**, Chapter 3), how should we communicate uncertainty when we do (**RQ2**, Chapter 4), and how could we tune models to match users' preferences with respect to uncertainty (**RQ3**, Chapter 5)? Through these questions, I have contributed to an understanding of what goes wrong when we communicate uncertainty poorly (by showing negative effects on trust in real users of weight scales), how we can communicate uncertainty better (by introducing the quantile dotplot visualization for communicating uncertainty in continuous measures), and how we can tune models to reflect users' error preferences in a lightweight way (by introducing my acceptability of error survey instrument). I take a holistic view of the design of everyday sensing and predictive systems, from the choices made in optimizing the underlying algorithms all the way to information presentation and how those design decisions interact to produce usable and understandable representations of uncertainty.

By considering these questions together, we can both design more effective representations of uncertainty *and* make more informed decisions about the algorithms we use: each choice affects the other and should be made in concert. In my work on end-user understanding of scales, it became clear that users' perceptions of the types of errors involved in that data have a strong effect on trust and acceptability. I believe there is much to be gained by more explicitly considering these types of errors in feedback design by more clearly reflecting an underlying model of uncertainty. I believe a better—more trustworthy—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 precise; greater gains may be had by pushing the state of the art in feedback.

This observation also strongly motivated my investigation into systematically estimating how acceptable users find the error of a system to be, as it seemed to me that only a heuristic answer to this question is normally given in novel sensing research. This persistent question—is my system accurate enough?—not only deserves a clearer answer, but represents stating point for improving user interfaces

and model selection. By expressing a model of acceptability of error in the domain language of users of machine learning algorithms, my approach allows us to easily adopt a classifier evaluation method that more closely matches users' perceptions of error than does the oft-used unweighted F-measure, while still being familiar to algorithm designers. At the same time, this method yields insight into how to build the application's feedback and whether further work on the classifier faces diminishing returns.

Greater adoption of these types of evaluation methods and more explicit consideration of uncertainty in the feedback of end-user predictive and sensing systems is becoming increasingly important. Estimation and prediction are more and more a part of our daily lives, and will only become more so: soon, smartphone and smartwatch applications for sensing heartrate, or blood sugar, or blood pressure, will be ubiquitous. If we truly want these applications to be ubiquitous—and not ubiquitously abandoned—we need to more effectively communicate the uncertainty in their estimates. Rather than waving error away, we should treat it as something to be integrated into the design of these systems from top to bottom—the only other option may be systems that don't make mistakes in the first place.

8. References

1. Antifakos, S., Schwaninger, A., and Schiele, B. Evaluating the Effects of Displaying Uncertainty in Context-Aware Applications. *Ubicomp '04*, (2004).
2. 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.
3. Barrowman, N.J. and Myers, R.A. Raindrop Plots: A New Way to Display Collections of Likelihoods and Distributions. The American Statistician 57, 4 (2003), 268–274.
4. Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4)
5. Bellotti, V., Back, M., Edwards, W.K., Grinter, R.E., Henderson, A., and Lopes, C. Making sense of sensing systems: five questions for designers and researchers. *chi '02*, (2002), 415–422.
6. Ben-Akiva, M., Lerman, S. (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. mit Press.
7. Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*.
8. bipm, iec, ifcc, ilac, iupac, iupap, iso, oiml. The international vocabulary of metrology—basic and general concepts and associated terms (vim3). jcgm 200:2012.
9. cdc Guidelines for Defining "Overweight" and "Obesity". http://www.cdc.gov/obesity/adult/defining.html
10. Centre for International Economics. (2001). Review of willingness-to-pay methodologies (pp. 1–48). Canberra.
11. Choe, E.K., Consolvo, S., Jung, J., Harrison, B., Patel, S.N., and Kientz, J. A. Investigating receptiveness to sensing and inference in the home using sensor proxies. *UbiComp '12*, (2012), 61.
12. Choo, H. and Franconeri, S.L. Enumeration of small collections violates Weber's law. Psychonomic bulletin & review 21, 1 (2014), 93–9.
13. 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.
14. Consolvo, S., Chen, M.Y., Everitt, K., and Landay, J.A. Conducting in situ evaluations for and with ubiquitous computing technologies. *hci* 22, (2007), 103–118.
15. Consolvo, S., McDonald, D.W., Toscos, T., *et al.* Activity sensing in the wild: a field trial of ubifit garden. *chi '08*, (2008), 1797–1806.

16. Cook, K.A. and Thomas, J.J. Illuminating the path: The research and development agenda for visual analytics. Pacific Northwest National Laboratory (pnnl), Richland, wa, 2005.
17. Correll, M. and Gleicher, M. Error bars considered harmful: Exploring alternate encodings for mean and error. ieee transactions on visualization and computer graphics 20, 12 (2014), 2142–2151.
18. Cumming, G. Inference by eye: reading the overlap of independent confidence intervals. Statistics in medicine 28, 2 (2009), 205–220.
19. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *mis quarterly* 13, 3 (1989), 319–340.
20. 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.
21. Domingos, P. MetaCost: A General Method for Making Classifiers Cost-Sensitive. *Proceedings of the fifth acm sigkdd international conference on Knowledge discovery and data mining* 55, (1999), 155–164.
22. 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.
23. Elkan, C. The Foundations of Cost-Sensitive Learning. *ijcai'01*, (2001), 973–978.
24. Feigelson, E.D. and Babu, G.J., eds. Statistical Challenges in Modern Astronomy. Springer New York, New York, ny, 1992.
25. Ferris, B., Watkins, K., & Borning, A. (2010). OneBusAway: results from providing real-time arrival information for public transit. *chi '10*, 1807–1816.
26. Finger, R. and Bisantz, A.M. Utilizing graphical formats to convey uncertainty in a decision-making task. Theoretical Issues in Ergonomics Science 3, 1 (2002), 1–25.
27. Froehlich, J., Larson, E., Campbell, T., Haggerty, C., Fogarty, J., and Patel, S.N. HydroSense: infrastructure-mediated single-point sensing of whole-home water activity. *Ubicomp '09*, (2009).
28. Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: a cross-cultural comparison. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 31(3), 444–57.
29. Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating Health Risks With Visual Aids. *Current Directions in Psychological Science*, 22(5), 392–399.
30. Gigerenzer, G. and Hoffrage, U. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102, 4 (1995), 684–704.
31. von Glasersfeld, E. Subitizing: The role of figural patterns in the development of numerical concepts. Archives de Psychologie 50, 194 (1982), 191–218.

32. Goldberg, R. and Hebbard, G. How accurate are hospital scales? Med J of Australia 194, 12 (2011), 665.
33. Gschwandtnei, T., Bogl, M., Federico, P., and Miksch, S. Visual Encodings of Temporal Uncertainty: A Comparative User Study. ieee Transactions on Visualization and Computer Graphics 22, 1 (2016), 539–548.
34. Gupta, S., Reynolds, M.S., and Patel, S.N. ElectriSense: single-point sensing using emi for electrical event detection and classification in the home. *Ubicomp '10*, (2010).
35. Haslam, D.W., James W.P. (2005). Obesity. Lancet 366 (9492): 1197–209.
36. Heer, J., Kong, N., and Agrawala, M. Sizing the horizon. Proceedings of the 27th international conference on Human factors in computing systems - chi 09, acm Press (2009), 1303.
37. Hoffrage, U. and Gigerenzer, G. Using natural frequencies to improve diagnostic inferences. Academic medicine : journal of the Association of American Medical Colleges 73, 5 (1998), 538–540.
38. Hullman, J., Resnick, P., and Adar, E. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. PloS one 10, 11 (2015).
39. Ibrekk, H. and Morgan, M.G. Graphical Communication of Uncertain Quantities to Nontechnical People. Risk Analysis 7, 4 (1987), 519–529.
40. Jackson, C.H. Displaying Uncertainty With Shading. The American Statistician 62, 4 (2008).
41. Johnson, B., & Slovic, P. (1995). Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. *Risk Analysis*, 15(4), 485–494.
42. Joslyn, S. and LeClerc, J. Decisions With Uncertainty: The Glass Half Full. Current Directions in Psychological Science 22, 4 (2013), 308–315.
43. jr, H. Redesigning the OneBusAway Mobile Experience. 2015. Master's thesis, Georgia Institute of Technology.
44. Jung, M.F., Sirkin, D., and Steinert, M. Displayed Uncertainty Improves Driving Experience and Behavior : The Case of Range Anxiety in an Electric Car. Proceedings of the 33rd Annual acm Conference on Human Factors in Computing Systems (chi '15), (2015), 2201–2210.
45. Kampstra, P. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. Journal of Statistical Software 28, code snippet 1 (2008), 1–9.
46. Kay, M., Choe, E.K., Shepherd, J., *et al.* Lullaby: a capture & access system for understanding the sleep environment. Ubicomp '12, (2012).

47. Kay, M., Kola, T., Hullman, J.R., and Munson, S.A. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. *CHI '16*, (2016).
48. Kay, M., Morris, D., schraefel, m. c., & Kientz, J. A. There's No Such Thing as Gaining a Pound: Reconsidering the Bathroom Scale User Interface. *Ubicomp '13*, (2013), 401–410.
49. Kay, M., Nelson, G.L., and Hekler, E.B. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. *CHI '16*, ACM Press (2016), 4521–4532.
50. Kay, M., Patel, S.N., and Kientz, J.A. How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. *CHI '15*, (2015), 347–356.
51. Kent, S. Words of Estimative Probability. Studies in Intelligence, (1964).
52. 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.
53. Koleva, B., Anastasi, R.O.B., Greenhalgh, C., *et al.* Expected, sensed, and desired: A framework for designing sensing-based interaction. *TOCHI* 12(1), (2005).
54. 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.
55. Kruschke, J.K. Bayesian data analysis. Wiley Interdisciplinary Reviews: Cognitive Science 1, 5 (2010), 658–676.
56. Kruschke, J. *Doing Bayesian Data Analysis*. Elsevier, (2011).
57. Larson, E.C., Lee, T., Liu, S., Rosenfeld, M., and Patel, S.N. Accurate and privacy preserving cough sensing using a low-cost microphone. *UbiComp '11*, (2011).
58. Leiss, W. (1996). Three phases in the evolution of risk communication practice. *The Annals of the American Academy of Political and Social Science* 545, 85–94.
59. Li, X., Wang, Y.-Y., and Acero, A. Learning query intent from regularized click graphs. *IGIR '08*, (2008), 339.
60. Lim, B.Y. and Dey, A.K. Assessing Demand for Intelligibility in Context-Aware Applications. *Ubicomp '09*, (2009), 195–204.
61. Lim, B.Y. and Dey, A.K. Investigating Intelligibility for Uncertain Context-Aware Applications. *Ubicomp '11*, (2011).
62. MacEachren, A. (1992). Visualizing uncertain information. *Cartographic Perspectives*, (13), 10–19.
63. MacKinlay, J. Automating the design of graphical presentations of relational information. ACM Transactions on Graphics 5, 2 (1986), 110–141.

64. McElreath, R. Statistical rethinking: A Bayesian course with examples in R and Stan. CRC Press, Boca Raton, FL, 2016.
65. Merino-Castello, A. (2003). Eliciting consumers preferences using stated preference discrete choice models: contingent ranking versus choice experiment. *UPF Economics and Business Working Paper*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=562982
66. 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.
67. Or, C.K.L. and Karsh, B.-T. A systematic review of patient acceptance of consumer health information technology. *JAMIA* 16, 4, 550–60.
68. Pak, S.S., Hutchinson, J.B., and Turk-Browne, N.B. Intuitive statistics from graphical representations of data. Journal of Vision 14, 10 (2014), 1361–1361.
69. Patel, S.N., Robertson, T., Kientz, J.A., Reynolds, M.S., and Abowd, G.D. At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line. *UbiComp '07*, (2007), 271–288.
70. Pavlou, P. Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. *IJEC* 7, 3 (2003), 69–103.
71. Pentland, A. and Choudhury, T. Face recognition for smart environments. *Computer*, February, (2000).
72. Popescu, M. and Li, Y. An acoustic fall detector system that uses sound height information to reduce the false alarm rate. *IEEE EMBS*, (2008), 4628–4631.
73. 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.
74. Potter, K., Kniss, J., Riesenfeld, R., and Johnson, C.R. Visualizing summary statistics and uncertainty. Computer Graphics Forum 29, 3 (2010), 823–832.
75. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2, 1 (2011), 37–63.
76. 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.
77. Rayson, P. and Garside, R. Comparing Corpora using Frequency Profiling. CompareCorpora 2000, 1–6.
78. Rigby, R.A. and Stasinopoulos, D.M. Using the Box–Cox t distribution in GAMLSS to model skewness and kurtosis. Statistical Modelling 6, 3 (2006), 209–229.
79. Van Rijsbergen, C.J. Evaluation. In *Information Retrieval*. Butterworth & Co., (1975), 95–132.

80. Rogers, Y., Connelly, K., Tedesco, L., *et al.* Why it's worth the hassle: The value of in-situ studies when designing UbiComp. *UbiComp '07*, (2007), 336–353.
81. Rukzio, E., Hamard, J., Noda, C., and Luca, A. De. Visualization of Uncertainty in Context Aware Mobile Applications. *MobileHCI '*06, (2006), 247–250.
82. 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.
83. Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., & Moorhead, R. J. (2009). A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *ieee Transactions on Visualization and Computer Graphics*, 15(6), 1209–18.
84. Scholtz, J. and Consolvo, S. Toward a framework for evaluating ubiquitous computing applications. *ieee Pervasive Computing* 3, 2 (2004), 82–88.
85. Silver, N. The Weatherman Is Not a Moron. *The New York Times*, (2012). http://www.nytimes.com/2012/09/09/magazine/the-weatherman-is-not-a-moron.html.
86. Smithson, M. and Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological methods 11, 1 (2006), 54–71.
87. Spiegelhalter, D.J. Surgical Audit: Statistical Lessons from Nightingale and Codman. Journal of the Royal Statistical Society. Series A (Statistics in Society) 162, (1999), 45–58.
88. Stan Development Team. Stan Modeling Language: User's Guide and Reference Manual. 2015.
89. Sweeting, H.N. Measurement and Definitions of Obesity In Childhood and Adolescence: A field guide for the uninitiated. Nutrition Journal 6, 32 (2007).
90. Taylor, B. N., & Kuyatt, C. E. (1994). Guidelines for Evaluating and Expressing the Uncertainty of nist Measurement Results, *nist Technical Note* 1297.
91. 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.
92. Tomlinson, J., Dyson, P. & Garratt, J. (2001). Student misconceptions of the language of error. U Chem Ed 5, 1–8.
93. Tufte, E. R. 2006. Beautiful Evidence.
94. Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
95. Van der Meulen, M., Logie, R. H., Freer, Y., Sykes, C., McIntosh, N., & Hunter, J. (2010). When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, 24(1), 77–89.

96. 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).
97. Varian, H. R. (2006). Revealed preference. In *Samuelsonian economics and the twenty-first century* (pp. 99–115).
98. Venkatesh, V. and Davis, F.D. A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management science* 46, 2 (2000).
99. Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F.D. User acceptance of information technology: Toward a unified view. *mis quarterly* 27, 3 (2003), 425–478.
100. Vrecko, D., Klos, A., and Langer, T. Impact of Presentation Format and Self-Reported Risk Aversion on Revealed Skewness Preferences. Decision Analysis 6, 2 (2009), 57–74.
101. 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).
102. 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.
103. Ward, A., Jones, A., and Hopper, A. A new location technique for the active office. *ieee Personal Communications*, October, (1997), 42–47.
104. Watkins, K. E., Ferris, B., Borning, A., Rutherford, G. S., & Layton, D. (2011). Where Is My Bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. *Transportation Research Part A: Policy and Practice*, 45(8), 839–848.
105. Wilkinson, L. Dot Plots. The American Statistician, (1999).
106. 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).
107. Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only anova Procedures. *chi '11*, (2011), 143–146.
108. Zuk, T. and Carpendale, S. Theoretical Analysis of Uncertainty Visualizations. spie-is&t Electronic Imaging 2006.

A. Appendix—Acceptability of error survey example

A.1 INTRODUCTION

This appendix gives an example acceptability of error survey as described in Chapter 5.

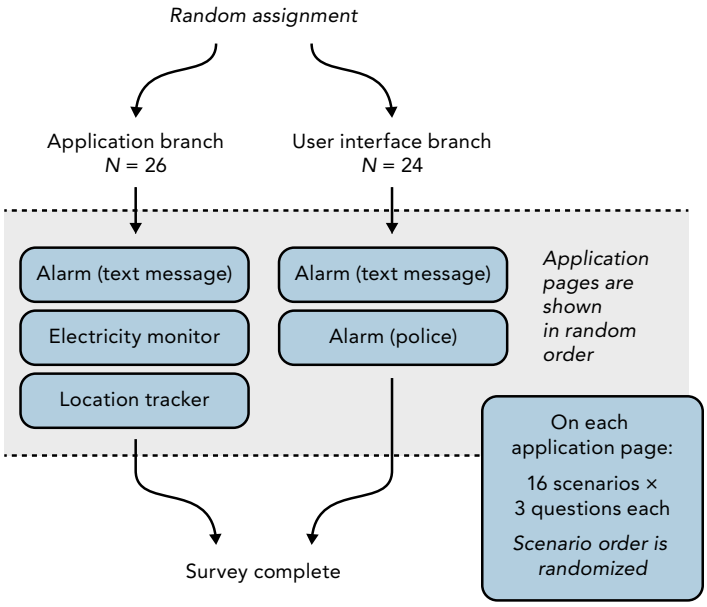
We are conducting a research study that aims to better understand the acceptability of accuracy in computer systems. We estimate this survey will take approximately 20-25 minutes to complete. Please answer each question as completely and honestly as you can. There is no risk to participating in this study. You may skip any question you do not wish to answer and withdraw from the study at any time.

At the end of the study, you will be given the opportunity to provide your email address to enter in a drawing to win a $50 Amazon.com gift card or one of 4 $25 Amazon.com gift cards (each participant has a chance to win at most 1 gift card per 1000 participants). Please note that you don't have to provide your email address if you wish to keep your response anonymous. However, all of the information will be confidential. We will record email addresses, if you provide them, and keep them in a list separate from and not connected to the data. To begin the survey, please click the "Next" button. By clicking next, you agree to participate in this study, that you understand you can withdraw from the survey at any time, that you should refrain from providing identifiable data in open-ended questions, and that you are at least 18 years of age. If you have any questions or concerns, please contact Matthew Kay (mjskay@uw.edu) or Julie Kientz (jkientz@uw.edu). We cannot ensure the confidentiality of any information sent by email.

A.2 SCENARIOS

The rest of this survey consists of a series of questions divided into three pages. Each page begins with a different scenario. **Please read the scenario at the top of each page before answering the questions on that page.**

[*Applications and scenarios in this section are randomized according to the diagram at right*]



A.2.1 Electricity monitor

In each of the questions in this section, we will ask you to read a brief description of the accuracy of **an electricity monitoring system for your home**. We will ask you to indicate how you feel about a system with the accuracy described. The accuracy of the system will be different for each question. The differences are indicated in bold.

Please imagine the following scenario:

Your residence has been outfitted with an intelligent electricity monitoring system. It is capable of keeping track of how often you use each of your appliances and how much electricity each appliance uses.

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **10 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **0 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **0 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer.

[*Likert item table:*

	Extremely unlikely	Quite unlikely	Slightly unlikely	Neither	Slightly likely	Quite likely	Extremely likely
I would find the accuracy of this system to be acceptable							
I would find this system to be useful							
If available to me now, I would begin using this system sometime in the next 6 months							

]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **10 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.

    - **0 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **2 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
    - **10 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **0 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **5 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
    - **10 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **0 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **10 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
    - **8 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **2 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **0 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **8 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **2 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **2 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **8 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **2 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **4 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **8 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **2 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **8 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **7 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **3 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.

- **0 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **7 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **3 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **1 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **7 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **3 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **4 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
  - **7 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
  - **3 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **7 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.

    - **5 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **0 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **1 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **3 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

Please imagine the following:

- **10 times** over a three month period, you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (correctly) reported that you used your clothes dryer.
    - **5 of the 10 times** that you used your clothes dryer, the system (incorrectly) reported that you used a different appliance.
- **5 other time(s)** over the same three month period, the system (incorrectly) reported that you used your clothes dryer even though you were actually using a different appliance.

[*Likert item table*]

If you have any questions or comments about this scenario, application, or the importance of accuracy in it, please enter them below.
[*Free-form text*]

A.2.2 Location tracker

In each of the questions in this section, we will ask you to read a brief description of the accuracy of **a location tracking system for your workplace**. We will ask you to indicate how you feel about a system with the accuracy described. The accuracy of the system will be different for each question. The differences are indicated in bold.

Please imagine the following scenario:

Your workplace has installed a mobile application on employees' cell phones that can estimate what room at work you or your coworkers are currently in. You can use it to locate a colleague or your supervisor when you are both present at work, for example, to have a quick meeting.

[*As in the Electricity application above, the following scenario is repeated over a space of hypothetical values of precision and recall, with bold text changed accordingly. For more details, see Chapter 5.*]

Please imagine the following:

- **10 times** over a three day period, your supervisor was in their office when you looked up their location.
  - **10 of the 10 times** that your supervisor was actually in their office, the system (correctly) reported that they were in their office.
  - **0 of the 10 times** that your supervisor was actually in their office, the system (incorrectly) reported that they were in a different room.
- **0 other time(s)** over the same three day period, when you looked up your supervisor's location, the system (incorrectly) reported that your supervisor was in their office.

[*Likert item table*]

If you have any questions or comments about this scenario, application, or the importance of accuracy in it, please enter them below.
[*Free-form text*]

A.2.3 Alarm (police)

In each of the questions in this section, we will ask you to read a brief description of the accuracy of **an alarm system for your home**. We will ask you to indicate how you feel about a system with the accuracy described. The accuracy of the system will be different for each question. The differences are indicated in bold.

Please imagine the following scenario:

Your residence has been outfitted with an intelligent alarm system that is capable of automatically recognizing household members when they enter, without any other interaction. For example, it does not require a password. When a stranger enters the house alone (someone that the system does not recognize), it calls the police.

[*As in the Electricity application above, the following scenario is repeated over a space of hypothetical values of precision and recall, with bold text changed accordingly. For more details, see Chapter 5.*]

Please imagine the following:

- **10 times** over a five year period, a stranger entered the house alone.
  - **10 of the 10 times** that this happened, the system (correctly) triggered the alarm and called the police.
  - **0 of the 10 times** that this happened, the system (incorrectly) did not trigger the alarm and did not call the police.
- **0 other time(s)** over the same five year period, the system (incorrectly) triggered the alarm.

[*Likert item table*]

If you have any questions or comments about this scenario, application, or the importance of accuracy in it, please enter them below.
[*Free-form text*]

A.2.4 Alarm (text message)

In each of the questions in this section, we will ask you to read a brief description of the accuracy of **an alarm system for your home**. We will ask you to indicate how you feel about a system with the accuracy described. The accuracy of the system will be different for each question. The differences are indicated in bold.

Please imagine the following scenario:

Your residence has been outfitted with an intelligent alarm system that is capable of automatically recognizing household members when they enter, without any other interaction. For example, it does not require a password. When a stranger enters the house alone (someone that the system does not recognize), it sends you a text message.

[*As in the Electricity application above, the following scenario is repeated over a space of hypothetical values of precision and recall, with bold text changed accordingly. For more details, see Chapter 5.*]

Please imagine the following:

- 10 times over a five year period, a stranger entered the house alone.
  - 10 of the 10 times that this happened, the system (correctly) triggered the alarm and sent you a text message.
  - 0 of the 10 times that this happened, the system (incorrectly) did not trigger the alarm and did not send you a text message.
- 0 other time(s) over the same five year period, the system (incorrectly) triggered the alarm.

[*Likert item table*]

If you have any questions or comments about this scenario, application, or the importance of accuracy in it, please enter them below.
[*Free-form text*]