How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy

Matthew Kay
Computer Science & Engineering | dub, University of Washington
mjskay@uw.edu

Shwetak N. Patel
Computer Science & Engineering | dub, University of Washington
shwetak@uw.edu

Julie A. Kientz
Human-Centered Design & Engineering | dub, University of Washington
jkientz@uw.edu

ABSTRACT
Many HCI and ubiquitous computing systems are characterized by two important properties: their output is *uncertain*—it has an associated accuracy that researchers attempt to optimize—and this uncertainty is *user-facing*—it directly affects the quality of the user experience. Novel classifiers are typically evaluated using measures like the F_1 score—but given an F-score of (e.g.) 0.85, how do we know whether this performance is good enough? Is this level of uncertainty actually tolerable to users of the intended application—and do people weight precision and recall equally? We set out to develop a survey instrument that can systematically answer such questions. We introduce a new measure, *acceptability of accuracy*, and show how to predict it based on measures of classifier accuracy. Out tool allows us to systematically select an objective function to optimize during classifier evaluation, but can also offer new insights into how to design feedback for user-facing classification systems (e.g., by combining a seemingly-low-performing classifier with appropriate feedback to make a highly usable system). It also reveals potential issues with the ubiquitous F_1-measure as applied to user-facing systems.

Author Keywords
Classifiers; accuracy; accuracy acceptability; inference; machine learning; sensors.

INTRODUCTION
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 [12] or fall detection [20], and using electrical noise and water pressure monitoring to track appliances' water and electricity use [9]. A common thread runs across these systems: they rely on inference, hence their output has *uncertainty*—it has an associated accuracy that researchers attempt to optimize—and this uncertainty is *user-facing*—it directly affects the quality of the user experience.

Consider an application that monitors energy usage of appliances 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. [17] 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 F_1 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? 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 Patel et al. [17] by Gupta et al. [9] improved mean 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, we have developed a model and method for predicting how acceptable users will find the accuracy of systems that use classifiers. The primary contributions of this paper center on connecting *evaluation of classifiers* to *acceptability of accuracy*, with the aim of predicting the latter on the basis of the former. These contributions are 1) formalizing the notion of acceptability of accuracy; 2) demonstrating the association between traditional measures of classifier accuracy and acceptability of accuracy (we investigate a class of

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

http://dx.doi.org/10.1145/2702123.2702603

weighted means of precision and recall that includes the ubiquitous F-measure and show how to use acceptability of accuracy to select which of measure to use when evaluating a classifier); and 3) devising and validating a simple survey tool developers of inference-based systems can use to help identify acceptable levels of classifier performance before expending the effort to build the systems.

As an example, imagine we are developing a smart alarm for the home that automatically identifies intruders and alerts the homeowner with a text message. While we have built a classifier for this problem, we are unsure what objective function to optimize: will our users value recall over precision, and if so, by how much? We generate a survey that describes the application and asks users how acceptable they find its accuracy to be in hypothetical scenarios with varying precision and recall (given a scenario description, the tool described in this paper generates the necessary scenarios). We deploy the survey to potential users then fit a model of acceptability of accuracy to the results. This model estimates α, a parameter from 0 to 1 describing the relative weight users place on precision versus recall: 0 means users value only recall, 1 only precision, and 0.5 each equally. The model estimates alpha at ~0.35 (some preference for recall over precision) and yields an objective function we can use to tune the classifier before we test it in a deployment. This increases the chance that our deployment is a success, as our classifier is now tuned to users' preferences for different types of errors in this application.

Our goal is not to impose further requirements for researchers to demonstrate that their good classifiers are actually good, though we believe it is possible to make such claims stronger through consideration of acceptability of accuracy. Instead, we aim to provide researchers with the tools to systematically make decisions about how to allocate resources: *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 "good enough".

In what follows, we outline our proposed survey instrument for assessing the acceptability of accuracy of hypothetical classifier-based applications. We describe a series of surveys in which we refined and validated our method. First, we employed this instrument to assess its face validity in four different applications drawn from the ubiquitous computing literature. Second, we deployed a refined version of the model and demonstrate its predicative validity in the domain of weather forecasting error, showing that we can estimate acceptability of accuracy to within one point on a 7-point Likert scale. We then discuss how we envision use of this tool in research and implications of this work on classifier evaluation when building systems.

RELATED WORK

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 [14,15]. The effects of intelligibility seem to be application-dependent: displaying uncertainty sometimes has positive [1] or negative [24] effects on task performance. In a study of several hypothetical context-aware systems, Lim and Dey 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 [15]. However, in the context of an inference-based system, it is not clear what components of accuracy contribute to assessments of "good enough." For example, it is well-established in information retrieval literature that the unweighted F_1 score is inadequate for many applications, since users may be more concerned (for example) with precision than recall [13,22]. Yet, we still commonly use F_1 score in evaluating classifiers in many user-facing applications. In this paper, we investigate the individual effects of precision and recall on the acceptability of accuracy in inference-based applications.

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, we use a variant of the Technology Acceptance Model (TAM) to validate our measure of acceptability of accuracy. TAM is a well-studied method for predicting technology acceptance, originally proposed for use in the workplace [6]. Since then, numerous variants of TAM have been proposed [27,28], and it has been applied to contexts outside the workplace, such as e-commerce [18] and consumer health technology [16]. 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 [6,18,27]. In this work, we adopt a variant of the TAM2 [27], which includes a construct called *output quality*—how well a system performs the tasks it is designed for—which we believe to be related to acceptability of accuracy 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, and several frameworks have been proposed [2,10,25]. These frameworks aim for a holistic evaluation, whereas we explicitly look toward a method for assessing the acceptability of accuracy. Others call for evaluating ubicomp technologies through in-situ deployment studies of built systems [23]. This can be a very useful method to assess the acceptability of accuracy, and studies of applications that use sensing have been able to evaluate the acceptability of accuracy of an already built system within the context of use (e.g., [5]). These deployments are very resource-intensive, however, and thus we aim to reduce the overhead of assessing the acceptability of accuracy before such systems are built. Finally, other researchers have proposed methods of formative assessment of ubicomp systems through the concepts of sensor proxies [3] and experience sampling [4], but these methods still require

in person interaction with participants, and do not provide explicit guidance on the acceptability of accuracy of inference systems. We believe our method can complement these existing approaches. In particular, by modeling acceptability of accuracy 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—we provide a model of acceptability of accuracy that is expressed in the domain language of the experts who build these systems.

ACCEPTABILITY OF ACCURACY SURVEY INSTRUMENT

We designed a scenario-based survey instrument to systematically examine the effects of differing classifier accuracies on user's 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:

> **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 [8]. 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, we 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).[1] Note that due to the definitions of recall and precision,

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

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

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

For each accuracy scenario, we ask three 7-point Likert-item questions from *extremely unlikely* to *extremely likely*. These questions correspond to *acceptability of accuracy* (which we introduce here), *perceived usefulness*, and *intent to use* (the latter two are adapted from the TAM [6,27])[2]:

- *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.*

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 accuracy. While we have selected particular levels of accuracy here, our scenario-generating code accepts any combinations of levels.

TESTS OF FACE VALIDITY

We intend our survey to be able to answer several questions about a given application. First, we aim to model *acceptability of accuracy* based on measures of classifier accuracy. To do that, we 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 accuracy.

Acceptability of accuracy as we define it is intended to correspond to a measure of *output quality* in TAM2 [27], which refers to how well a system performs the tasks it is designed for (distinct from how useful someone finds those tasks to be) and has been shown to correlate with *perceived usefulness* [27]. This leads to our first test of validity:

- **T1:** *Acceptability of accuracy* and *perceived usefulness* should be highly correlated.

Further, per TAM [6,27]:

- **T2**: *Perceived usefulness* and *intent 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

---

[1] The use of frequencies necessitates some rounding, so some scenarios have only approximately this precision/recall.

[2] 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 we omitted it.

errors will vary between applications, and our instrument should uncover this; thus:

- **T3**: Our instrument should be sensitive to *application*: classifiers with similar accuracy for different applications may have different *acceptability of accuracy*.

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 [22], 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 our method to be sensitive to such differences in situations where the costs of errors differ. Thus, our fourth test:

- **T4**: When classifiers with similar accuracy for *the same* application have different levels of user burden for false positives, our test should be sensitive to this, and reflect it as a different weighting of *precision* versus *recall*.

STUDY 1: ASSESSING FACE VALIDITY

We deployed our instrument in a survey with four different hypothetical applications inspired by applications found in the ubicomp literature [9,19,29]. This variety was intended to allow us to validate T3. 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 on the alarm application are meant to explore two possible extremes of feedback: a relatively low-burden modality (text messages) and a very high-burden modality (calls to the police). These allow us to validate T4.

Survey structure for data collection

Due to the length of the survey (each application has 16 scenarios, with 3 questions per scenario), we split the survey into two branches (see Figure 1). Each participant is randomly assigned to one of two branches: the *application branch* and the *user interface branch*, corresponding to T3 and T4. 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.

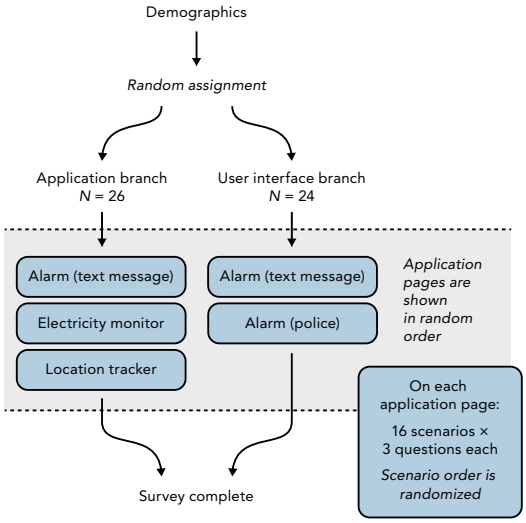

Figure 1. Survey 1 structure. Each application page (blue box) corresponds to an instance of our acceptability of accuracy survey instrument applied to a different application.

Participants

Participants were recruited via word-of-mouth, distribution on university mailing lists, and advertisements on Facebook. We had 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. We used randomization of scenario order to account for this so that each application still received an adequate number of participants. We had 50% female participants, and 50% of each branch was female.

Model of acceptability of accuracy

To analyze acceptability of accuracy, we posited that acceptability of accuracy may be predicted based on some measure of classifier accuracy. In particular, because precision and recall in these applications are visible to the user, we concentrated on measures based on them. We did not consider measures that involve true negatives, as it is not clear in our 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, we first consider the weighted F-measure, which is equivalent to a weighted harmonic mean of precision and

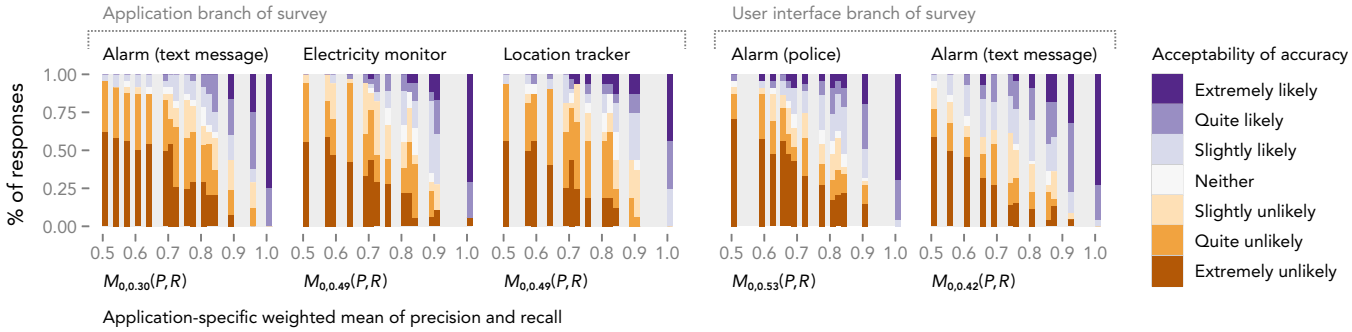

Figure 2. Acceptability of accuracy from the survey results plotted against each application's weighted geometric mean of precision and recall from our model. Optimizing this mean has the effect of also optimizing an application's acceptability of accuracy.

recall. We 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 [22]), or the F_{β}-measure where $\alpha = 1/(1+\beta^2)$; thus $M_{-1,0.5}$ is the familiar F_1 score. $M_{0,\alpha}$, the geometric mean, is also known as the G-measure [21]. We consider this larger class of measures so that we have a systematic way to ask both whether harmonic mean (i.e., F measure) corresponds most closely to how people judge acceptability of accuracy for these applications (by determining p) and so that we can estimate whether for a given application, people value precision or recall more highly (by determining α for that application).

We conducted a mixed-effects Bayesian logistic regression[3] of *acceptability of accuracy* against three different weighted power means of precision and recall (harmonic, geometric, and arithmetic). Our 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 we consider $acceptability_{i,j,k} = 1$ when a participant rates the acceptability of accuracy for that scenario as *Slightly likely* or higher and 0 otherwise. U_k is the random effect for participant k. We used the following uninformed priors:

$$\begin{aligned} \beta_{0,i}, \beta_{1,i} &\sim \text{Normal}(0,1000) \\ \alpha_i &\sim \text{Uniform}(0,1) \\ U_k &\sim \text{Normal}(0, 1/\tau) \\ \tau &\sim \text{Gamma}(0.001,0.001) \\ p+2 &\sim \text{Categorical}(1/3, 1/3, 1/3) \end{aligned}$$

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.

We 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: 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 our 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 our model (and hopefully in the future, others') can be used to inform prior distributions in future work. We adopt Kruschke's [11] approach to Bayesian experimental statistics. In particular, we examine 95% highest-density intervals (HDIs) of posterior distributions to estimate credible differences between parameters (as opposed to an NHST approach of a 0.05 p-value threshold on the distribution of a test statistic).[5]

RESULTS

The posterior distribution of p allows us to estimate which measure best approximates acceptability of accuracy. For these applications, p = 0 (geometric mean) is most credible (P(p = 0) = .81), suggesting the G-measure may more close-

[3] While we 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.

[5] Where possible we also ran similar more traditional NHST models and saw similar effects.

ly correspond to users' estimations of accuracy here. We do note 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 2 plots the proportion of people who rated the acceptability of accuracy at each level against the weighted geometric mean for that application derived from our model. Higher weighted mean is clearly associated with greater acceptability of accuracy. We break down the rest of our results based on the validity questions outlined above.

- **T3**: Our instrument should be sensitive to *application*: Classifiers with similar accuracy for different applications may have different *acceptability of accuracy*.

Confirmed. See Figure 3A: 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* ($\alpha_i - \alpha_j = -.19$, 95% HDI: [−.30,−.09]) and between *alarm (text message)* and *location tracker* ($\alpha_i - \alpha_j = -.19$, 95% HDI: [−.32,−.07]) were credible on a 95% HDI, suggesting that our tool can be sensitive to differences in preferences between applications. β_1 for all applications on both branches was also credibly different from 0. While it varied, $\beta_1/100$ 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. While their posterior distributions of α were similar, *electricity monitor* and *location tracker* had credible differences in β_1 ($\beta_{1,i} - \beta_{1,j} = 6.6$, 95% HDI: [.3,12.6]), again showing the sensitivity of the model.

- **T4**: When classifiers with similar accuracy for *the same* application have different levels of user burden for false positives, our test should be sensitive to this, and reflect it as a different weighting of *precision* versus *recall*.

Confirmed. See Figure 3C: In the user interface branch of the survey, the alarm scenario had $\alpha = 0.53$ for police compared to the much lower $\alpha = 0.42$ for alarm with text message[6], and these differences were credible on 95% HDI ($\alpha_i - \alpha_j = -.12$, 95% HDI: [−.22,−.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

[6] While estimates of α for alarm (text message) differed between branches, a model combining both branches (Figure 3B) yields a more precise estimate of α that is consistent with the separate estimates (within their 95% HDIs) and which has all of the same credible differences with other αs as described above.

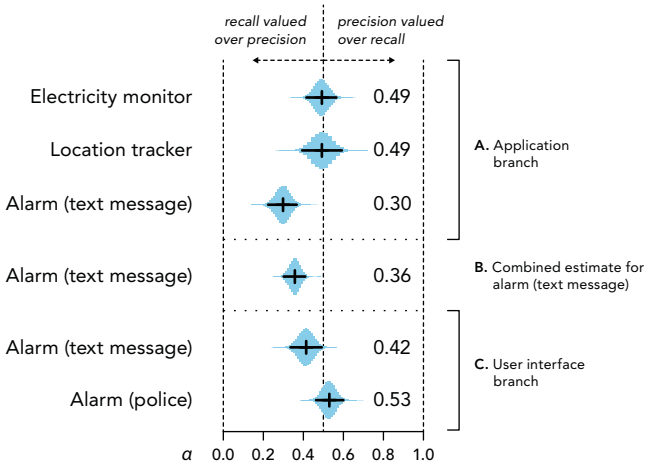

Figure 3. Posterior distributions of α for both branches of the survey. Mean and 95% HDI are indicated. Note the sensitivity of our model to different preferences of precision versus recall between applications (A) and for different feedback types (C). (B) shows a more precise estimate of α for alarm (text message) from a model combining both branches of the survey.

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.

- **T1:** *Acceptability of accuracy* 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 our inclusion of *acceptability of accuracy* as a measure of output quality in the TAM.

- **T2**: *Perceived usefulness* and *intention to use* should be highly correlated.

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

STUDY 2: ASSESSING PREDICTIVE VALIDITY

To assess the predictive validity of our tool, we conducted a survey of 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. We obtained ground truth data of precipitation predictions from various weather forecasters in Seattle, WA, USA for the period of Sept 1, 2013–Aug 31, 2014 (http://www.forecastwatch.com/). We focused on one city, as people in deferent climates may have different preferences for precipitation accuracy. This survey had two parts:

Part 1: Existing acceptability (ground truth). We asked participants to specify which weather forecasting apps and websites they currently use and to rate each on *acceptability of accuracy* of precipitation prediction over the *last 30 days*

and the *last year*. We also included *usefulness, ease of use,* and *frequency of use* questions for validating against the TAM. We 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 our survey tool, these questions do not specify the accuracy of the systems in question to participants. However, since we have the ground truth of the predictive accuracy of these systems over both time periods in question, we can model these responses in a similar manner to our hypothetical accuracy survey without the caveat that people are responding to hypothetical levels of accuracy.

Part 2: Hypothetical acceptability (our survey tool). We randomly selected one application or website from Part 1 that the participant currently uses and generated a variant of our survey instrument for that application or website. We asked them 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 accuracy*, *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, we specified TP, FN, and FP (and additionally TN, as we had 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." We used the same precision and recall levels as before, but instead of giving all 16 scenarios to each participant, we gave each participant a random subset of 8 scenarios to reduce survey length.

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), we used an ordinal regression instead of a binomial regression. Our model assumes the same latent variable representing acceptability is associated with the ordinal ratings of acceptability of accuracy in Part 1 and in Part 2.

Besides the use of ordinal instead of binomial regression, we also added two additional parameters. We 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 accuracy when answering about a system they have experienced. We 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:

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

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

We use leave-one-participant-out cross-validation to assess the predictive validity of our model. In each fold, we 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). We 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 system is known, but the real-world acceptability of future (possibly better) systems is unknown. This scenario might arise (e.g.) 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.

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

Results

Our model had a cross-validated mean absolute error (MAE)—the mean difference between the predicted and actual acceptability in points on the Likert scale—of 0.93, suggesting our predictions are generally within one point of the actual value on the scale. This is promising, though we note 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 in our 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]).

We found some evidence of *wet bias* [26] in participants' preferences: the estimated α was 0.461 (95% HDI: [0.382,0.545]) with 82.3% of the distribution lying below

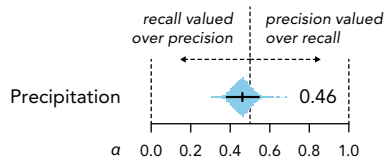

Figure 5. Posterior distributions of α for precipitation prediction. Mean and 95% HDI are indicated. Note the evidence of wet bias: 82% of the distribution lies below 0.5.

0.5 (see Figure 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. We expect the prevalence of this bias to vary by climate, so make no claims to generalizability beyond the city we tested in. We note that we 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 our 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* ($\rho = 0.95$, $p < 0.001$). We also saw significant correlations between *acceptability of accuracy* 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 concerns like usability and convenience have additional salience over accuracy). Notably, we did not see significant correlations between *acceptability of accuracy* and *ease of use* in the ground truth ($\rho = 0.13$, $p = 0.13$), but did see moderate correlations between *ease of use* and *usefulness* ($\rho = 0.55$, $p < 0.001$)—as predicted by the TAM—suggesting that *acceptability of accuracy* is a separate construct from *ease of use* and is more related to *output quality* and *usefulness*, as we had predicted.

DISCUSSION AND IMPLICATIONS

In this section, we provide a discussion and implications for the survey instrument we have built for estimating acceptability of accuracy and its potential uses. Our 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.

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 accuracy ground truth, our 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), we believe that this estimated objective function will correlate with real-world acceptability of accuracy more closely than (say) F_1 score. More broadly, we believe researchers should consider whether F_1-measure truly matches their evaluation goals before employing it on user-facing systems.

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, we 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 our 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 our instrument). This gives us 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 accuracy of a system, it may be premature to rule out a weak—but adequately-performing—classifier without investigating acceptability of accuracy 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 we believe our 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.

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 [9] or water [7] 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 spark arguments between parents and children about TV-time quotas. We could more systematically investigate these intuitions by fitting a model of acceptability of accuracy to each of these applications. This would allow us to decide if our classifier is adequate for each use case.

Predicting future acceptability and setting targets

Given actual use of a classifier with known accuracy, acceptability ratings of that accuracy, and results from our survey instrument, we can estimate the relationship between

hypothetical and actual acceptability (as with the weather data above). In this case, we can actually use hypothetical ratings to estimate the acceptability of accuracy for future classifiers that are more accurate than existing ones, and use this model to set targets for desired accuracy or to identify when we have reached a point of diminishing returns.

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 monitor); such systems wait until they have built a good personalized model. But how long should training last? First impressions made by poor predictions are likely to sour users on a system. Given a model of acceptability of accuracy for an application, one could set a desired threshold of acceptability (e.g., as a percent of the user base), and use this to determine when the system should switch from training to prediction.

Expanding to other application domains

Thus far we have examined four specific application domains: electrical appliance detection, person location within an office, home alarms, and precipitation prediction. We chose applications we felt would be broadly applicable to a general audience and that we could use 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 we did not test them here). We 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 our approach is designed to determine.

To facilitate adoption, we provide code for generating the survey based on desired precision and recall levels and fitting the model at: https://github.com/mjskay/acceptability-of-accuracy. We 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.

Recommendations on applying the survey to research

Our experience in conducting this research leads us to make several recommendations for researchers hoping to apply a similar approach to their own applications and classifiers. We recommend presenting each user with at most 8 accuracy scenarios (as we did for the weather application), as we received feedback that the original survey (with 16 scenarios) was a bit long. We also recommend including at most two applications at a time, as our 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 we used written scenarios in our example, 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 could be made a single, small task.

LIMITATIONS AND FUTURE WORK

While we believe that our approach can be useful to help give researchers an easy to use method for assessing acceptable accuracy levels for a given classifier and interface, there are some limitations. First, the models are typically application-specific. However, as described in the discussion, we believe that it is straightforward to use existing classifiers in a domain to derive a model for that domain, allowing prediction of acceptability of accuracy of future classifiers. A good next step for this 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 accuracy can be predicted using our acceptance of accuracy survey (similar to how we validated the precipitation prediction model). We also believe that model estimates from previous, similar applications can inform future models (and here, our Bayesian approach can facilitate this). Finally, as an initial test case for our 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 or to incorporate a broader set of classifier evaluation measures (c.f. [21]).

CONCLUSION

This work was motivated by a persistent question in HCI and ubiquitous computing research with end-user feedback based on classifiers: is my classifier good enough? We introduced a new measure, *acceptability of accuracy* and developed and validated a survey instrument that connects classifier evaluation to acceptability of accuracy. By expressing our model in the domain language of classifier designers, our approach allows us to easily adopt an evaluation method that more closely matches users' perceptions of accuracy than does the oft-used unweighted F-measure. 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. We advocate for greater adoption of these types of evaluation methods in user-facing classifiers through the use of a community database of models of acceptability in HCI application domains.

ACKNOWLEDGEMENTS

We thank Sean Munson and Eun Kyoung Choe for their valuable feedback on this work and Cynthia Matuszek for her particular insight into the problems discussed herein. This work was funded in part by the Intel Science and Technology Center for Pervasive Computing (ISTC-PC).

REFERENCES

1. Antifakos, S., Schwaninger, A., and Schiele, B. Evaluating the Effects of Displaying Uncertainty in Context-Aware Applications. *UbiComp '04*, (2004).
2. 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*, 1 (2002), 415–422.
3. 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.
4. Consolvo, S., Chen, M.Y., Everitt, K., and Landay, J.A. Conducting in situ evaluations for and with ubiquitous computing technologies. *Int J Hum-Comput Int 22*, (2007), 103–118.
5. 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.
6. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly 13*, 3 (1989), 319–340.
7. 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).
8. Gigerenzer, G. and Hoffrage, U. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review 102*, 4 (1995), 684–704.
9. 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), 139–148.
10. Koleva, B., Anastasi, R.O.B., Greenhalgh, C., et al. Expected, sensed, and desired: A framework for designing sensing-based interaction. *ACM Transactions on Computer-Human Interaction 12*, 1 (2005), 3–30.
11. Kruschke, J.K. *Doing Bayesian Data Analysis.* Elsevier Inc., 2011.
12. 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).
13. Li, X., Wang, Y.-Y., and Acero, A. Learning query intent from regularized click graphs. *SIGIR '08*, (2008), 339.
14. Lim, B.Y. and Dey, A.K. Assessing Demand for Intelligibility in Context-Aware Applications. *Ubicomp '09*, (2009), 195–204.
15. Lim, B.Y. and Dey, A.K. Investigating Intelligibility for Uncertain Context-Aware Applications. *Ubicomp '11*, (2011), 415–424.
16. Or, C.K.L. and Karsh, B.-T. A systematic review of patient acceptance of consumer health information technology. *JAMIA 16*, 4, 550–60.
17. 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).
18. Pavlou, P. Consumer acceptance of electronic commerce: integrating trust and risk with the technology acceptance model. *Int J Electron Comm 7*, 3 (2003).
19. Pentland, A. and Choudhury, T. Face recognition for smart environments. *Computer*, February (2000).
20. 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.
21. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J of Mach Lear Tech 2*, 1 (2011), 37–63.
22. Van Rijsbergen, C.J. Evaluation. In *Information Retrieval*. Butterworth & Co., 1975, 95–132.
23. 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.
24. Rukzio, E., Hamard, J., Noda, C., and Luca, A. De. Visualization of Uncertainty in Context Aware Mobile Applications. (2006), 247–250.
25. Scholtz, J. and Consolvo, S. Toward a framework for evaluating ubiquitous computing applications. *IEEE Pervasive Computing 3*, 2 (2004), 82–88.
26. Silver, N. The Weatherman Is Not a Moron. *The New York Times*, 2012. www.nytimes.com/2012/09/09/magazine/the-weatherman-is-not-a-moron.html.
27. Venkatesh, V. and Davis, F.D. A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Manage Sci 46*, 2 (2000), 186–204.
28. 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.
29. Ward, A., Jones, A., and Hopper, A. A new location technique for the active office. *IEEE Personal Communications*, October (1997), 42–47.