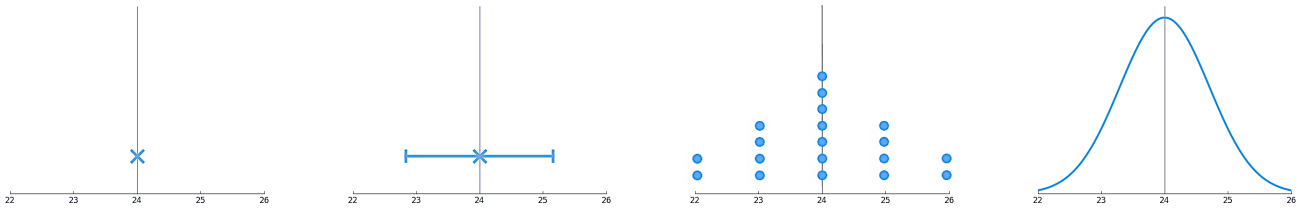
Uncertainty Visualization Influences how Humans Aggregate Discrepant Information

Miriam Greis[1], **Aditi Joshi**[2], **Ken Singer**[1], **Albrecht Schmidt**[1,3], **Tonja Machulla**[1]
[1]University of Stuttgart, Stuttgart, Germany, {firstname.lastname}@vis.uni-stuttgart.de
[2]Olin College of Engineering, Needham MA, US, {firstname.lastname}@students.olin.edu
[3]LMU Munich, Munich, Germany, {firstname.lastname}@ifi.lmu.de



(a) Point Estimate (b) Confidence Interval (c) Dotplot (d) Probability Density Function

Figure 1. We used four different visualizations with an increasing amount of represented uncertainty information (from left to right) to understand how uncertainty information influences humans' choice of internal models for information aggregation.

ABSTRACT

The number of sensors in our surroundings that provide the same information steadily increases. Since sensing is prone to errors, sensors may disagree. For example, a GPS-based tracker on the phone and a sensor on the bike wheel may provide discrepant estimates on traveled distance. This poses a user dilemma, namely how to reconcile the conflicting information into one estimate. We investigated whether visualizing the uncertainty associated with sensor measurements improves the quality of users' inference. We tested four visualizations with increasingly detailed representation of uncertainty. Our study repeatedly presented two sensor measurements with varying degrees of inconsistency to participants who indicated their best guess of the "true" value. We found that uncertainty information improves users' estimates, especially if sensors differ largely in their associated variability. Improvements were larger for information-rich visualizations. Based on our findings, we provide an interactive tool to select the optimal visualization for displaying conflicting information.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3174079

Author Keywords

Uncertainty; visualization; information aggregation; conflicting information.

INTRODUCTION

We are increasingly surrounded by electronic devices, many of which contain various types of sensors. Applications use these device sensors to track various personal and environmental attributes and present them to users. These data are uncertain since the measurement process is corrupted by noise, the sensors may be wrongly calibrated or the algorithms use thresholds to determine sensor values. As the number of similar devices increases, the amount of conflicting and potentially confusing information likewise increases: if several sensors are redundant—i.e, they measure the same attribute—the results may not always converge on one and the same value. Conflicting values can be automatically aggregated; however, our prior work showed that users prefer to have access to individual data sources even if automatic aggregation is available [14]. Here, we investigate how humans reconcile such inconsistent sensor information and whether different visualizations of the uncertainty associated with sensor measurements improve the quality of inference. The goal is to provide guidelines for the design of interfaces that present and summarize sensor data from multiple sources.

As an example for a situation when sensor data may disagree, we introduce Marc who—as a passionate biker—is always eager to keep track of how much he has biked. He has a mobile phone that uses GPS tracking while he is biking as well as a sensor on his bike that tracks the distance based on the number of revolutions the wheel has turned. At the end

of the day, Marc checks an application on his phone, which can synchronize with the phone's GPS tracker and the wheel revolutions sensor. One of the two sensors may tell him that he has biked 13.9 km while the other recorded a distance of 14.7 km. Both of these sensor results are uncertain and Marc has to decide on how he makes sense of the conflicting information. A similar situation may arise if Marc only owned one sensor and used it to assess the distance of the same route on several separate occasions—the sensor may tell him that he biked 13.9 km one day, 14.5 km another day, and 14.2 km on the third day.

There are different ways how Marc could reconcile the sensor data to figure out what he believes the true value to be. For example, he could always believe his GPS tracker because the recorded values may have a smaller variance and thus a higher reliability. Alternatively, he could combine his two sensor measurements into one value, for instance by choosing a value in-between the two measurements. It is likely that his strategy to aggregate the sensor data will differ depending on the visualization of the sensor information and the amount of uncertainty communicated by it.

The goal of our work is to understand how different uncertainty visualizations (see Figure 1) influence humans in aggregating conflicting probabilistic data. There are several ways of aggregating probabilistic data, such as taking a simple average, using a maximum likelihood estimator, or a winner-takes-all model [8]. Budescu et al. [4] found that people combine information by averaging information. However, in many situations, this approach does not lead to the optimal solution. Here, we conducted a user study with 16 participants, to investigate whether different visualizations of sensor data and their associated uncertainty influences the users' selection of a model for data aggregation. In the study, participants were shown two visualization of sensor data from two hypothetical sensors next to each other and asked what they believed the true value to be. We used four different visualizations encoding different amounts of uncertainty information.

The three contributions aligned to the goal of our research are as follows:

1. We show that humans **aggregate conflicting information differently if they have access to uncertainty information**. They do not build a simple average, but rather weigh each source of information taking its reliability into account.

2. We show that **the amount of uncertainty information included in a visualization influences the internal models** humans use for combining information.

3. Based on our findings, we provide **an interactive tool** that aids interface designers in the selection of uncertainty visualizations best suited to improve users' inference.

Our results demonstrate that humans interpret conflicting information differently depending on the amount of presented uncertainty information. By providing an interactive tool for designers and other researchers to explore our results, we allow to easily apply them to real-world application scenarios and support quick comparison for future research purposes.

BACKGROUND

Uncertainty is an inherent part of many data sets, but often the data is presented as if it was perfectly precise and accurate [7]. The consequences are that users of a system might think that the system is more reliable than it truly is. If they make a wrong decision based on the false impression of reliability, they can lose their trust in the system [1]. Therefore, the communication and visualization of uncertainty is not to be neglected. However, it is associated with many research challenges. For example, Brodlie et al. [3] outline that uncertainty is a complex topic itself and that many different visualizations can be used to communicate uncertainty. We are therefore interested in how such different visualizations influence humans in interpreting uncertain sensor measurements.

Definitions and Classifications of Uncertainty

The term "uncertainty" is widely used throughout different disciplines, albeit often inconsistently. Therefore, many domain-specific definitions exist. In addition, the differentiation from related concepts such as data quality, reliability, accuracy, and error often remains underdefined [25]. The disagreement between domains and researchers can be illustrated by contrasting two example definitions. Pang et al. [27] define uncertainty "to include statistical variation or spread, error and differences, minimum-maximum range values, noisy, or missing data." They choose a very open definition to include all possible cases. In contrast, Gershon [11] developed a very narrow definition of uncertainty defining it as one of several sources of imperfect knowledge, namely a case where the accuracy of a fact is not known to the user. For the purposes of the present work, we refer to uncertainty as the error of measurements obtained via sensors.

In addition to the plethora of definitions, many typologies and taxonomies exist which are built on different concepts. They either focus on the domain such as geospatial or medical information [39, 31], visualization approaches [27, 29], the level of the data [36], or the source of uncertainty [17].

Inconsistent Sensor Measurements

Measurements obtained via two different devices can differ from one another. The reason for this is that any measurement process suffers from inherent variability. This variability leads to measurement error—that is, there is a deviation of the measurement from the true value of the quantity being measured. There are two types of error: random error and systematic error [38].

All measurement systems, including sensor systems, are subject to *random error*. Random error is also called Gaussian noise, because the error magnitudes follow a Gaussian distribution. Such nonsystematic fluctuations can be introduced at any level of the measurement process: sensors introduce imprecisions when registering input, the signal transmission through electronic circuits is hampered by imperfections, analog-to-digital conversions and other operations on the signal are imprecise and last but not least the human operator makes mistakes when reading off the measurement result (e.g., when measuring with sub-millimeter precision using a standard ruler). The variability of the random error directly determines

the *reliability* or precision of a measurement device—the lower the noise, the higher the reliability. A highly reliable device delivers very similar measurements if the conditions surrounding the measurement are kept consistent.

A second source of inconsistency between two measurement devices is *systematic error*, or bias. A biased device will generate measurements that differ by a consistent amount from the true value and thus from measurements performed with an unbiased device. Systematic errors can result from poor initial calibration or instead accumulate over time, as is the case for measurements using a gyroscope. The magnitude of the bias determines the *accuracy* of a device—perfectly accurate measurements are unbiased.

Random and systematic errors affect measurements independently of each other. Measurements can be reliable but imprecise, unreliable but precise, neither reliable nor precise, or both.

Importantly, when two devices provide diverging measurements, knowledge about the type and magnitude of measurement error improves the decision about what the true value is. In the present work, we assume a situation where the user has no possibility of assessing the presence of bias (e.g., with the help of the third device) and also does not have a reason to believe that one of the two measurement devices is more accurate than the other. In this case, the user should ideally treat any difference between the measurements as resulting from random error. In Bayesian terminology, we assume that the prior distribution is uniform. In this situation, what strategies are there to reconcile the measurements and are some strategies better than others?

Reconciling Inconsistent Measurements

A simple strategy to arrive at a single value is to pick one of the two measurements. This could be in the form of a winner-takes-all strategy, i.e., the user arbitrarily chooses to always disregard readings from one device. For instance, the user could choose to only consult his smartphone app with regards to the kilometers biked and ignore the sensor on the wheel. However, this defies the purpose of having access to multiple sources of information. Alternatively, on each instance, the user picks one of the devices at random, akin to flipping a coin. Both of these strategies are not optimal as they disregard part of the information available.

Another strategy is to take the arithmetic mean of the two sensor measurements. If the devices are indeed unbiased, this improves on the coin-flip strategy in that the user's estimate on average will be closer to the true value. To illustrate, a user repeatedly bikes 10 km and uses one unbiased device to measure this distance. The measured values will vary around 10 km and the mean magnitude of this error can be quantified by computing the standard deviation SD_{error} of the error. Since the user always chooses the sensor value as her estimate, her estimate will have the same standard deviation $SD_{estimate}$ as SD_{error}. In contrast, if the user has access to two unbiased devices with equal reliability (i.e., the devices have the same SD_{error}) and computes the arithmetic mean, the result will on average differ from the true value by $SD_{estimate} = \frac{SD_{error}}{\sqrt{2}}$. The user's combined estimate is in fact more reliable than the sensor measurements.

Lastly, what if the two sensors do not have equal reliability? The solution is to allow the more reliable device to have a stronger influence on the user's estimate. This can be achieved by computing a weighted average of the two measurements [38]:

$$X_{est.} = \frac{w_A \, Measurement_A + w_B \, Measurement_B}{w_A + w_B} \quad (1)$$

where the weight of each measurement are determined by the measuring device's reciprocal variance:

$$w_A = \frac{1}{Var_A} \quad and \quad w_B = \frac{1}{Var_B} \quad (2)$$

This solution is statistically optimal in that it maximizes the possible reliability of the user's estimate. In fact, it has been shown that humans often rely on weighted averaging when combining redundant information from different sensory modalities [2, 9] or from within the same modality [19].

RELATED WORK

Morss et al. [26] conducted a nation-wide survey in the US, in which they found that weather forecast users are aware of the uncertainty in forecasts and that users prefer to actually be informed about uncertainty. In the following, we present related work about the communication and visualization of uncertainty including perspectives from different disciplines such as psychology, visualization, and Human-Computer Interaction (HCI).

Textual Communication of Uncertainty

Uncertainty can be communicated in various ways. Verbal and numerical expressions are the basic forms of communicating probabilistic information to the general public. Both approaches have disadvantages. Humans may interpret verbal expressions such as "low uncertainty" or "almost certain" very differently [41]. They assign different probabilities for the same terms [5]. Numerical expressions, however, could as well be misinterpreted as humans in general have difficulties to answer even easy probability questions [24] or to understand the correct meaning of the probability of rain [12]. Psychological studies also demonstrated that humans are prone to biases and suboptimal heuristics when judging uncertain information in textual form [40].

Uncertainty Visualization

An alternative to textual communication of uncertainty are visual representations. Uncertainty visualization is a widely explored research topic. In the visualization community, many different visualizations have been explored in depth, such as glyphs [27, 43], line graphs [37], box plots [28], or bar charts [6]. Other strands of work focused on the visualization of uncertainty for the general public: e.g., Ibrekk et al. [18] compared different visualizations for weather forecasts and came to the conclusion that a probability density function in combination with a cumulative probability density function is the best option to communicate uncertainty to the general

public. Although there is a huge body of work on uncertainty visualization, it is far from resolved when to use which visualization as most studies focused on a very specific context.

Recently, the visualization of uncertainty has also gained attention in the HCI community [15] in the context of different application scenarios such as body weight measurements [23], data analysis [10], public transport predictions [22], genome data [35], or range anxiety in electric cars [21]. The vast majority of studies showed that adding uncertainty information has many advantages such as increasing trust and user experience (for a counterexample see [34]).

Decision-Making under Uncertainty

A multitude of studies in the field of psychology showed that providing uncertainty information has a positive influence on decision-making. In a study conducted by Roulston et al. [32], participants made better decisions as managers of a road salting company when standard errors were presented in addition to a point estimate. In a very similar study, Joslyn et al. [20] showed that concrete decision aids are only successful if uncertainty information are provided along with them. In other application scenarios, such as weather forecasting [33] or flood forecasting [30], similar results were observed.

Implications

Related work shows that the communication of uncertainty is an important topic and recently gaining attention in HCI. Humans are aware that data is uncertain and prefer to have uncertainty information presented. They also make better decisions if exposed to uncertainty information. However, there is still a need for better understanding when the use of different visualizations is beneficial, how they influence humans' internal models of how the data has been generated, and how inferences from the data are drawn.

In our study, we build on existing work when selecting uncertainty visualizations suited to communicate errors of sensor measurements. We can further contribute to the body of knowledge by providing concrete insights in how humans aggregate conflicting information, when uncertainty information is provided in the form of different visualizations. In many application fields in HCI, there may be contradicting information from different sources (such as contradicting sensor information), which have to be displayed to a user. Knowing how users interpret different visualizations of two data points and what models they inherently use to aggregate them can support researchers and designer to find more optimal ways of presenting their data.

RESEARCH QUESTIONS

Our main goal is to understand how people aggregate sensor information if interfaces provide them with discrepant information from two different sensors. Specifically, we pose four concrete research questions (RQ1 to RQ4), each with associated hypotheses.

RQ1: Does information about the uncertainty of sensor data influence how people aggregate inconsistent sensor measurements? Interfaces can show data with different degrees of detail. They can either show the two most likely sensor values without any uncertainty information, or they can use existing approaches to visualize uncertainty. We assume that visualizing uncertainty (no matter in which way) changes how people aggregate sensor information as they have additional information to make a judgment. They will use the additional information to adjust their judgment. This leads to *H1a: People use a simple average if no uncertainty information is presented.* and *H1b: People use more elaborate models to aggregate information if uncertainty information is presented.*

RQ2: If uncertainty is visualized, do people take the reliability of sensors into account? Specifically, do people make statistically optimal decisions? We assume that depending on the reliability of the presented sensor measurements, humans will combine the information from separate sources differently. For two measurements with the same reliability, we assume that humans build the average as examined by Budescu et al. [4]. For different reliabilities, we assume that users will not use a simple average, but rather weigh the presented information in a statistically optimal fashion. This leads to *H2a: People use a simple average if sensors have equal reliability.* and *H2b: People use a weighted average, as given by equation 1, if two measurements with different reliabilities are presented.*

RQ3: Are some visualizations better suited to improve decisions than others? The findings from Greis et al. [13] and Kay et al. [22] suggest that a visualization with detailed aggregated uncertainty information may be most effective, e.g., in the form of a histogram or a dotplot. This leads to *H3: People make better decisions (i.e., chose estimates closer to statistical optimality) if a visualization with detailed aggregated uncertainty information is used.*

RQ4: Do people adjust their strategy for data aggregation according to the inconsistency between two sensor measurements? If the inconsistency between two sensor measurements is much larger than the noise associated with each sensor, it is very likely that at least one of the sensors is biased. In this situation, aggregating the measurements into one value may lead to worse results than choosing one of the measurements, namely the less biased one. However, what if people have no way of assessing, which sensor is biased? Then they can either 1) ignore the potential presence of bias and use a weighted average as described above, or 2) use a coin-flip strategy to chose one of the measurements at random, or 3) always chose the sensor measurement with either the lower or the higher associated reliability. Of these two alternatives, the latter is more likely since people might assume that devices that are more reliable are also more accurate. This leads to *H4: People change their aggregation strategy if the sensor measurements disagree too much.*

METHODS

We conducted a user study in the laboratory to investigate our research questions in a controlled environment. In the following, we present the experimental design, task, stimuli, and the procedure for our experiment.

Experimental Design, Task and Stimuli

Our stimuli differed in terms of which visualization was used to communicate the reliability of the sensor measurements, how far the two sensor measurements diverged from one another and the relative difference in reliability between sensors. Specifically, we used a 4 x 5 x 3 within-subject design with three independent variables: *Visualization* (with the four levels point estimate, confidence interval, dotplot, probability density function), *Delta Sigma* (with the five levels 50%-50%, 40%-60%, 30%-70%, 20%-80%, 10%-90%), and *Inconsistency* (with the three levels no inconsistency, small inconsistency, large inconsistency). We generated randomly sampled values for means and SDs for each condition. In the following, we describe these factors as well as the stimuli for each of the factor levels.

Visualizations

We decided to focus on graphical representations of uncertainty to increase the comparability between our conditions. We used four visualizations with increasing amount of uncertainty information, as described by Greis et al. [16]. Specifically, we used a simple point estimate visualization (see Figure 1a) as a visualization with zero uncertainty information. As a first visualization of aggregated uncertainty information, we used a confidence interval of the size of the underlying distribution's variance (see Figure 1b) around the point estimate. For the aggregated detailed uncertainty information, we used a dotplot (as suggest by Kay et al. [22], see Figure 1c). Each dotplot contained 19-20 dots as the small number supports subitizing. The number of bins was determined by the SD and ranged from 3 to 13. For the full amount of uncertainty information, we used a traditional probability density function (henceforth, PDF), as depicted in Figure 1d.

Delta Sigma

We used five levels of difference in the relative reliability/variance between the two sensors. We call this factor Delta Sigma. For each pair of sensor measurements, we chose the associated sensor variances such that weight w_1 resulting from equation 2 increased in steps of 0.1 from 0.5 to 0.9 and w_2 decreased from 0.5 to 0.1 (for both weights after normalizing). For easier reference, we will denote these different levels as the 50%-50%, 40%-60%, 30%-70%, 20%-80%, and 10%-90% Delta Sigma level. In the 50%-50% level, both sensors have equal variability. In contrast, in the 10%-90% level, one sensor is very reliable and has low variability—it is therefore assigned a high weight—while the other sensor is very unreliable and is assigned a low weight. Figure 2c shows an example of two distributions with unequal variances.

Sensor Inconsistency

We used three different levels of inconsistency between the two sensor measurements, as depicted in Figure 2. Figure 2a shows two measurements where the presented distributions have zero inconsistency and the best aggregated estimate is therefore identical to the two measurements. Figure 2b shows two measurements with a small inconsistency and Figure 2c shows two measurements with a large inconsistency. The small and large inconsistencies equaled on average 1.9 and 5 times the size of the standard deviation of the distribution used in the 50%-50% conditions, respectively.

Stimuli and Task

The experiment consisted of 800 trials. During each trial, participants were presented with two visualizations of the same type, each depicting one hypothetical sensor measurement and the variability associated with the measurement (except in the case of the point estimate visualization, which does not provide uncertainty information). The absolute value of the first measurement X_1 was randomly drawn from the range of 12 and 6755. The second value X_2 was obtained by adding one of the three levels of inconsistency to X_1. The variances for the two values were obtained from $Var(X_1) = 5w_1$ and $Var(X_2) = 5 - 5w_1$.

The visualizations were presented one above the other. A slider was shown in-between the two visualizations (see of Figure 2a, 2b and 2c for an example of the interface). It allowed participants to select what they believed to be the "true value" of the quantity that the sensors were measuring. To finish a task, participants had to click a button to continue to the next pair of measurements. The task did not explicitly state what factors where measured or what type of sensors were used. We assume that such descriptions could have created a bias for participants as they might not trust a specific sensor due to their personal experience or knowledge. For each visualization, there were 200 trials, 40 for each Delta Sigma level. For each of these levels, participants experienced the zero inconsistency condition 8 times, and the other two inconsistency conditions 16 times each. The presentation location of a single visualized source (top-bottom, left-right) was randomized but not entered into the analysis as an additional variable since we had no a-priori hypothesis how location could influence decisions.

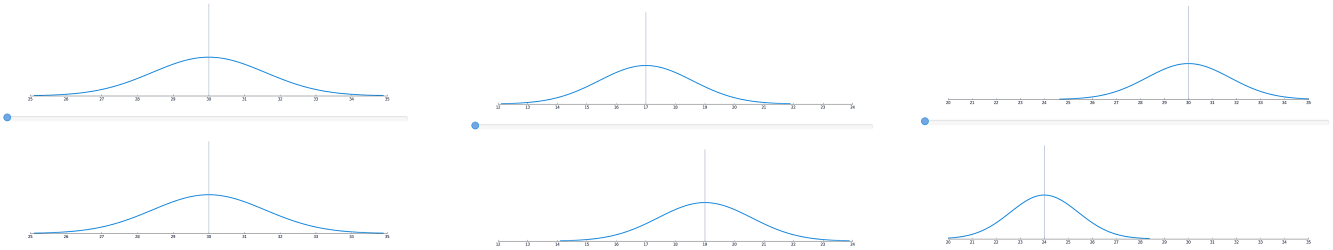
All 200 trials within the same visualization were presented in one block; the presentation order was balanced across participants using a Latin square design. Within each block, the presentation order of stimuli from different inconsistency and different Delta Sigma levels was randomized. However, we balanced the placement of the visualizations, i.e., whether X_1 was placed to the left or right of the combined average and whether it was displayed as the top or the bottom visualization.

Participants

We recruited 16 voluntary participants (8 male, mean age of 24.6 with a $SD = 3.4$). Most of the participants were university students.

Procedure

At the beginning of the study, participants filled in a basic demographic questionnaire before the experimenter explained the general task. Afterwards, we introduced participants to the topic of unreliable sensor measurements providing examples similar to the example in the introduction of this paper. Participants then completed 200 tasks for each visualization.



(a) Zero Inconsistency (b) Small Inconsistency (c) Large Inconsistency

Figure 2. We used three different levels of inconsistency between measurements in our experiment. The figure shows examples for each level.

Effect/Interaction	df	df.res	F	p
Visualization	3	12113	1201.969	$< .001$
Delta Sigma	4	12113	692.652	$< .001$
Inconsistency	2	12113	787.455	$< .001$
Visualization:Delta Sigma	12	12113	170.183	$< .001$
Visualization:Inconsistency	6	12113	215.389	$< .001$
Delta Sigma:Inconsistency	8	12113	175.496	$< .001$
Visual.:Delta Sigma:Inconsistency	24	12113	54.775	$< .001$

Table 1. Results of the linear mixed-effects model analysis on the aligned-rank transformed data for the three main factors visualization, Delta Sigma, and inconsistency and the random factor participant.

Before starting a new round of 200 tasks, participants got a short introduction on what the visualization depicts and one trial to get used to the new visualization. We additionally encouraged participants to take small breaks before starting with a new round of 200 tasks so as to not become fatigued during the study.

RESULTS

First, we calculated participants' estimation error on each trial. For this, we obtained the weights that participants assigned to each of the two presented sensor measurements from their chosen combined estimate X_{chosen}, using $\hat{w}_1 = \frac{X_{chosen} - X_2}{X_1 - X_2}$ and $\hat{w}_2 = 1 - \hat{w}_1$. We then calculated the difference between the predicted and the chosen weights for the sensor measurement with the higher reliability. For example, a participant saw two visualizations that were generated with a Delta Sigma level of 20%-80%. He then chose a combined estimate corresponding to a Delta Sigma level of 32%-68%. We calculate the difference for the higher weighted estimate, which here results in a deviation of the participant's estimate of 12% or 0.12 from the optimal choice.

To analyze whether the estimation error differed across conditions, we performed a linear mixed-effects model analysis on the aligned-rank transformed data [42] as the data was not normally distributed. We used the three fixed factors *visualization*, *Delta Sigma*, and *inconsistency* and the random factor *participant*. For all significant main effects, we conducted post hoc pair-wise comparisons of the factor levels. For significant two-way interactions, we conducted further post hoc tests in the form of differences of differences.

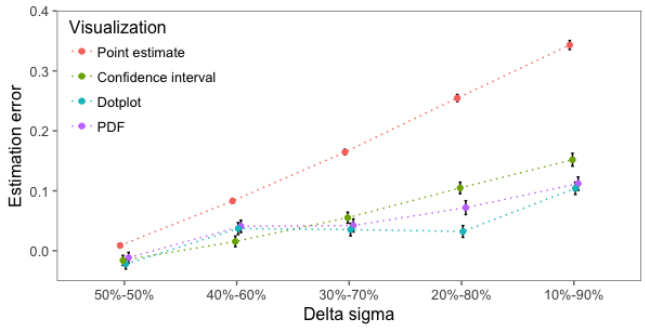
The analysis revealed that all three main effects for the visualization, the Delta Sigma, and the inconsistency are significant. However, the three two-way interactions and the three-way interactions are significant, as well, which might impact the interpretation of the main effects. Table 1 shows the results of the mixed-effects analysis. We will discuss these findings in more detail in the following. Results of post-hoc tests for interactions are available in the supplementary material. Due to the large number of possible comparisons between different conditions (most of which are statistically significant), we will focus the discussion on interesting observations that can be visually verified from Figures 3 and 4.

Visualization

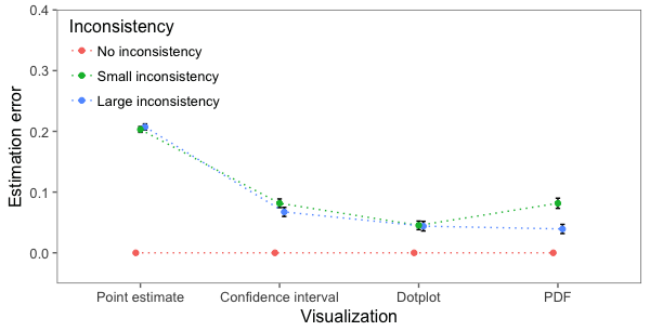
The post hoc tests revealed that the estimation error differed significantly between all pair-wise combinations of visualizations, with the exception of the combination of dotplot and the PDF. The average estimation error was highest for the point estimate ($M = 0.17, SD = 0.18$), followed by the confidence interval ($M = 0.06, SD = 0.24$), the PDF ($M = 0.05, SD = 0.26$), and the dotplot ($M = 0.04, SD = 0.24$). Notably, the estimation error for the latter three visualizations is only about one third the magnitude of the error in the point estimate conditions. These results support our hypothesis *H1b* that people aggregate information differently if uncertainty information is present in the visualization than if it is not present. However, our hypothesis *H3* cannot be accepted as the dotplot did not generally perform better than the PDF. Overall, however, participants performed closer to optimal when uncertainty information was included.

Delta Sigma

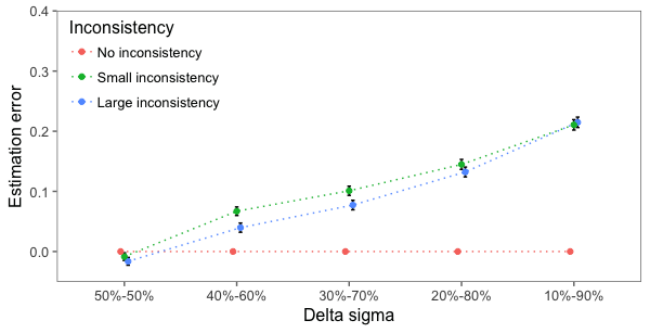
The post hoc tests revealed that the estimation error differed significantly for all pair-wise combinations of levels. The error was smallest for the 50%-50% Delta Sigma ($M = 0.01, SD = 0.19$). This supports our hypothesis *H2a* that people use a simple average if two information with equal reliability are presented. Further, the estimation error increases as the sensors differ more and more in their reliabilities from each other: the error is smallest for the 40%-60% level ($M = 0.04, SD = 0.22$), followed by the 30%-70% level ($M = 0.07, SD = 0.23$), the 20%-80% level ($M = 0.12, SD = 0.25$), and largest for the 10%-90% level ($M = 0.18, SD = 0.26$). This means that, although we find support for the hypothesis *H2b* that people use a weighted average when values with different reliabilities are presented, the chosen average value is not optimal, in particular for large differences in reliability.



(a) Mean estimation errors for each of the four visualizations as a function of the reliability difference Delta Sigma.



(b) Mean estimation errors for each of the three inconsistency levels as a function of the visualization.



(c) Mean estimation errors for each of the four visualizations as a function of the reliability difference Delta Sigma.

Figure 3. Participants' mean estimation errors for combinations of two variables; error bars represent standard errors of the mean.

Inconsistency

The post hoc tests revealed that the estimation error differs significantly between the three inconsistency levels. For the no inconsistency level, there was no error ($M = 0.00, SD = 0.00$). This supports our hypothesis *H4a* that people choose the best estimate if information do not disagree, which is not surprising. In the case that the measurements disagree, people on average chose a higher Delta Sigma for smaller inconsistencies ($M = 0.10, SD = 0.25$) than for the larger inconsistencies ($M = 0.09, SD = 0.26$). This agrees with our hypothesis *H4b* as the inconsistency seems to have an influence on how the measurements are aggregated.

Interaction between Visualization and Delta Sigma

Figure 3a shows the estimation error for the four visualizations as a function of the reliability differences between the two distributions. The general picture does not contradict the main effects of the factors Delta Sigma and visualization: the error increases with an increasing difference in sensor reliability and the point estimate performs worse than any of the three visualizations with associated uncertainty. An interesting feature of this interaction effect is that for the most extreme differences (20%-80%, 10%-90%), the dotplot and the PDF improve participants' judgments compared to the confidence interval.

Interaction between Visualization and Inconsistency

Figure 3b shows the estimation error for the three inconsistency levels as a function of the visualizations. Generally, the interaction is mostly in agreement with the conclusions drawn from the main effects: the error is smaller for visualizations with uncertainty and there is some indication that the error is larger for small sensor inconsistencies than for large inconsistencies, especially in the case of the PDF.

Interaction between Delta Sigma and Inconsistency

Figure 3c shows the estimation error for the three inconsistency conditions as a function of the reliability difference Delta Sigma between distributions. Again, the data generally supports the previous observations: the error increases with increasing difference between sensor reliabilities and the error is larger for smaller sensor inconsistencies. This last relationship appears to mostly result from the central three Delta Sigma levels. More insight is provided by the three-way interaction.

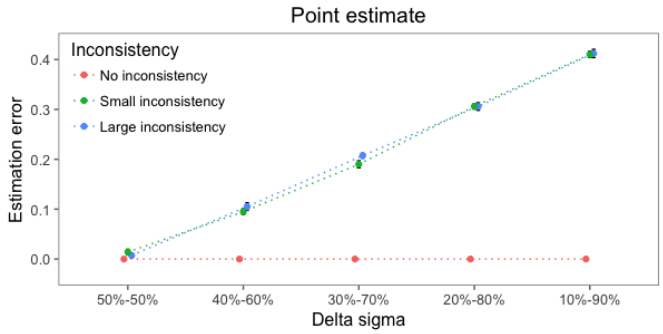
Three-way Interaction

Figure 4 shows the estimation error for the three inconsistency levels as a function of the difference in sensor reliability, separately for each of the four visualizations. This most clearly illustrates the following points: 1) when no information about sensor reliability is provided (see Figure 4a - point estimate), people chose the average value (this provides support for *H1a*); 2) across all visualizations, the estimation error increases with the increasing difference in sensor reliabilities; 3) this increase is less severe if uncertainty information is added to the visualization; 4) dotplot and PDF are more successful in reducing the error, in particular for large difference in sensor reliabilities; 5) the error is mostly larger for smaller inconsistencies, in particular in case of the PDF.

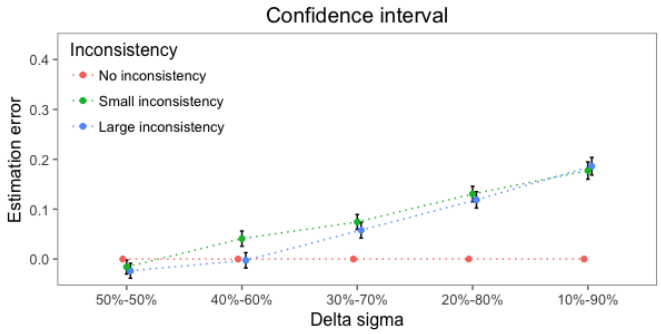
DISCUSSION

Our results show that the presentation of uncertainty information indeed changes how people aggregate conflicting data into one value. Without uncertainty information, people average the data as already suggested by Budescu et al. [4]. In contrast, if people are presented with information regarding the reliability of data, they use weighted averaging to arrive at a single value. This second strategy actually results in better estimates of the "true" value if neither of the sensors providing the data is biased.

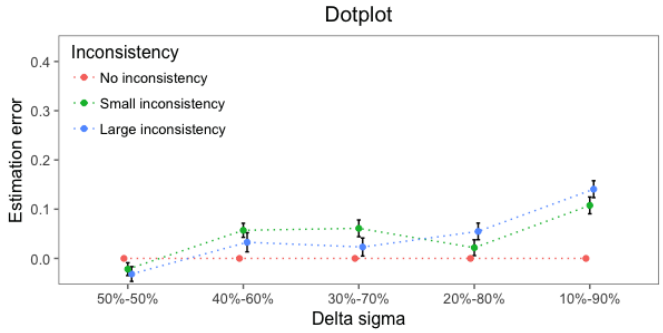
We further find that uncertainty visualizations with greater detail, such as the dotplot or the PDF, result in a better weighting



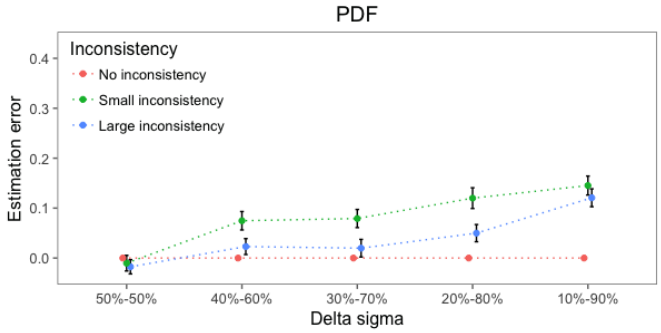
(a) Mean estimation errors for the point estimate for each of the three inconsistency levels as a function of the reliability difference Delta Sigma.



(b) Mean estimation errors for the confidence interval for each of the three inconsistency levels as a function of the visualization.



(c) Mean estimation errors for the dotplot for each of the three inconsistency levels as a function of the reliability difference Delta Sigma.



(d) Mean estimation errors for the PDF for each of the three inconsistency levels as a function of the reliability difference Delta Sigma.

Figure 4. Participants' mean estimation error separate for each visualization; error bars represent standard errors of the mean.

of the sensor data than a confidence interval. This is interesting since the confidence interval visualization contains all the information necessary to compute the optimally weighted average—that is, it shows the sensor measurements and their associated variances. One possible explanation for this is that it may be more difficult to correctly assess the relative magnitudes of the two variances from the confidence interval. For example people might be better at determining that the data from one sensor is three times as variable as the data from the other sensor when they see a dotplot covering a two-dimensional area rather than when they see a one-dimensional line as is the case for the confidence interval. While interesting, testing this possible explanation is beyond the scope of the present work. To summarize our results concerning the visualization of uncertainty, the data suggests that for maximal improvement in aggregating two inconsistent sensor measurement, people should be supplied with more complex visualizations such as dotplot and PDF.

By how much the aggregation process is improved through the use of uncertainty visualization also depends on the relative difference in sensor reliabilities as well as the magnitude of the inconsistency between the sensor measurements. We will discuss our findings in more detail in the following.

First and most obvious, for equally reliable sensors the value predicted by the weighted averaging strategy is the same as calculating the simple average. In this case, there is no advantage to show uncertainty information as the point estimate performs just as well as the other visualizations. However, as soon as sensor measurements diverge in reliability from one another, showing uncertainty information improves people's judgments. This improvement is quite dramatic: the error is more than halved in all unequal of the Delta Sigma levels where the sensors differ in their reliability.

However, there is still some residual error—that is, people's judgments are not optimal in a statistical sense. Deviation from the optimal choice of weights increases with increasing difference in reliability between the sensors. The direction of the error is such that participants are more conservative—their estimate is shifted towards the simple average (i.e., away from the measurement with the higher reliability). That is to say, people underestimate the need for choosing extreme weight differences. For example, in the case of 10%-90% weightings they might chose to weigh the sensor values with 25%-75%, instead. The reason for this could indeed be computational in nature: that is, an increasing underestimation of the relative size difference between the two variances. For example, in the 20%-80% weighting level one of the variances is 4 times the magnitude of the other but might be perceived to be only 3 times larger. For the next larger level of 10%-90%, one of the variances is 9 times larger than the other but might be perceived to be only 7 times larger. This increasing underestimation might result in an increasing estimation error. Another and

simpler reason could be that participants had a behavioral bias during the experiment: namely, to input their estimate on the slider away from the extreme points and more towards the middle section between the two distributions. Again, these explanations are offered as hypothetical interpretations of the data; in-depth differentiation between these options is beyond the scope of the current work and also does not relate to the main research questions. In sum, while the estimation error is largest for large differences in sensor reliabilities regardless of the visualization, these conditions also profit the most in terms of reducing absolute error as compared to not providing any uncertainty information at all.

Regarding the inconsistency between sensors measurements, there are two notable findings. First, participants always selected the best estimate when there is no inconsistency between sensor data, independent of weighting and visualization. This is, of course, trivial. Nevertheless, it is notable in so far that we have to conclude that when sensor measurements are very similar (in units of standard deviation of the more reliable sensor), there is no need to use a complex visualization as the point estimate leads to equivalent performance.

The second interesting finding is that participants' error is mostly larger for small as compared to large inconsistencies. There are several possible explanations for this.

First, the finding may be artefactual. The error is computed as a relative quantity. For this reason, a 1 mm deviation on the input slider when entering the estimate translates into a larger deviation of the chosen weight from the predicted weight for small absolute distances on the screen vs. large absolute distances. If this was the case the error should always be larger for small as compared to large distances. However, we find that this relationship is reversed for two of the extremer weighting levels in the dotplot (see Figure 4c). We can therefore reject this explanation.

A second possible explanation is that smaller inconsistencies may invoke a stronger behavioral bias in the direction of the simple average. This might be related to some properties of the visualization, such as a higher degree of overlap of the distributions.

Third, participants might have changed their strategy for obtaining a single estimate as sensor measurements became more inconsistent. As previously discussed, this makes sense because the larger the inconsistency relative to the sensor noise, the more likely it is that the difference partially results from systematic error of bias of one or both sensors. In the presence of bias, weighted averaging no longer leads to optimal results. In the present study, participants might have assumed that the less reliable measurement is also less accurate—that is, that it is biased—and shifted their estimates away from it. This would counteract the effect of underestimated unequal weighting and result in a smaller error. This explanation opens interesting possibilities for further research, such as assessing when people believe sensor measurements to differ due to systematic rather than random errors and how this affects decision making based on inconsistent data from multiple sources. In sum, our findings show that inference changes as sensor measurements become more inconsistent and highlight the need for further research.

Limitations

The present work models one basic aspect of sensor measurements—namely, corruption by Gaussian noise. We started our investigation with this type of error because for this case a formal, well-established model exists of how decisions can be optimized. The model generates quantitative predictions that serve as ground truth.

However, realistic sensor measurements are likely to also reflect other sources of error (e.g., the previously discussed bias), which may result in more complex error distributions (e.g., slanted, heavy-tailed, or non-continuous). Obtaining ground truth in such situations is more difficult and requires the development of novel mathematical formalizations of optimal decision making under uncertainty. While this was beyond the scope of the present study, the problem has to be addressed eventually so that visualizations of measurement variability can be tailored to support decision making rather than lead to erroneous conclusions.

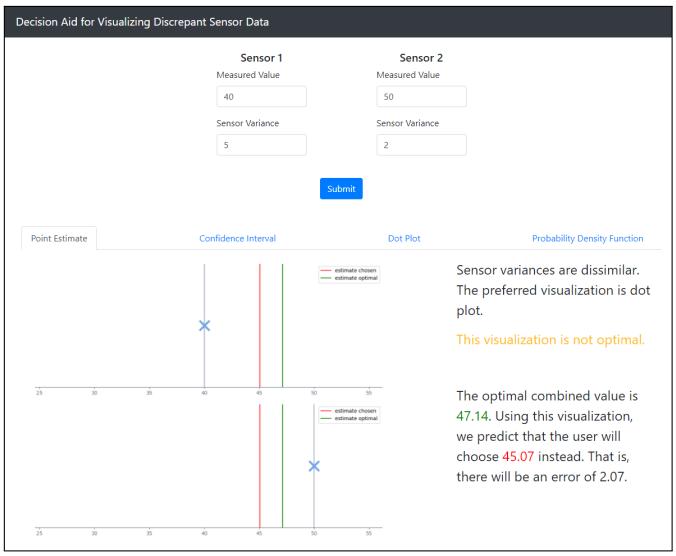
Another problem in realistic application scenarios consists in obtaining parameters for the generation of the visualizations—e.g., a mean and a standard deviation for the symmetrical confidence interval in Figure 1b. If there are no specs available on a sensor's reliability (e.g., from the manufacturer), the process of translating raw measurements into a representative visualization such as a continuous pdf will require some considerations (e.g., whether and how to capture skew, how to deal with sparse data etc.).

DESIGN RECOMMENDATIONS

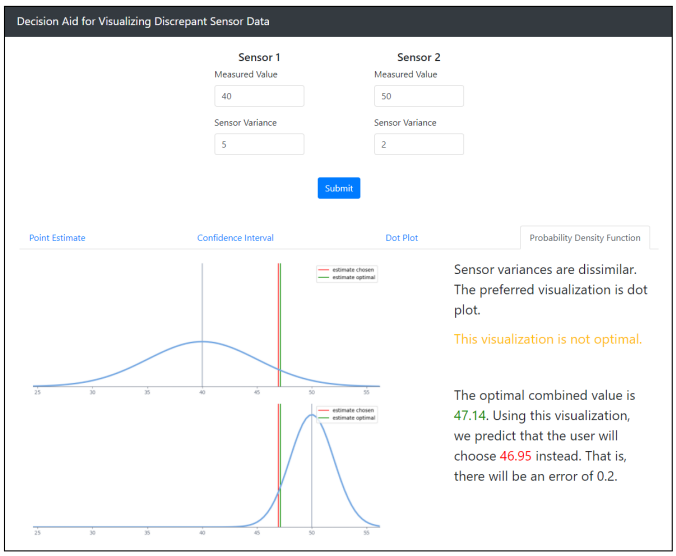
In the following section, we present recommendations for the choice of visualizations when designing interfaces that gather and present information from different sensors, all of which assess the same attribute. We believe such interfaces will increase in number as we are surrounded by more and more electronic devices that monitor our activity, health, interactions with other people, the status of devices and objects that belong to us like smart cars or homes, or environmental and societal conditions of interest.

We base our recommendations on the results of our study and to some degree on findings from previous research. We attempt to strike a balance between specificity and consistency of the interface—that is, between suggesting the visualization that has led to the best decision given a certain condition and keeping the visual characteristics of the interface as similar as possible across different instances of use.

From these considerations, we suggest a consistent use of either the dotplot or the PDF visualization. Both show the largest decreases in error across many conditions. Related work suggested a preference for the dotplot [22]. If the designer is comfortable with the interface switching between several visualizations from use to use, we suggest using simple point estimate visualizations under specific conditions, namely when either the sensor measurements or the reliabilities associated with each measurement are very similar. Measurements



(a) Interface for the design aid for the point estimate.



(b) Interface for the design aid for the PDF.

Figure 5. Screenshots of our design aid tool that shows our results for sensor measurements and variances entered by the user.

can be regarded similar if they differ by less than 0.5 standard deviations in terms of the variability of the more reliable sensor. Reliabilities can be considered similar if the normalized weighting pair assigned to the measurements by equation is less discrepant than 40%-60%. Under these circumstances, the user's aggregated estimate is not improved by providing a detailed visualization of uncertainty. However, even then seeing the inconsistency in relation to the sensor noise might help the user: namely by alerting him to a possible bias in one of the sensors.

For easier access of our results and recommendations, we implemented an interactive tool (see Figure 5). It allows the visual exploration of conflicting data from two sensors using the four visualizations studied in the current work. For each visualization, it provides the optimal combined value, a prediction of the expected error as well as our suggestion, which visualization to use for the particular data set. The tool can be downloaded on GitHub: https://github.com/hcilab-org/InformationAggregation.

CONCLUSION

In this paper, we conducted a study to understand how uncertainty visualizations influence humans' choice of internal models of information aggregation. We showed that presenting uncertainty information improves users' inference when the inconsistency between two measurements increases and when the measurements' associated reliabilities are dissimilar. We recommend to use a point estimate visualization only for measurements with little inconsistency or non-diverging reliabilities. The higher the inconsistency or difference in reliability, the more humans benefit from uncertainty information. Based on the interpretation of our results, we recommend to use a dot plot to communicate uncertainty information for conflicting information.

To better illustrate our results and make them easily explorable for designers and other researchers, we implemented and contribute an interactive tool that allows to visualize conflicting information. In addition to the visualizations, it shows the optimal combined value and a prediction of what users might actually choose as the combined value. We hope that our research will inspire research in other application areas of HCI to understand whether the results are applicable to communicating uncertain data in other contexts than sensor measurements. As we choose sensor measurements as a general scenario, we are confident that the results are applicable in other areas.

In future work, we hope to extend the interactive tool to support more different types of visualizations. Further, we are interested in understanding how the difference between systematic and random error and larger inconsistencies than tested in this experiment influence humans' inference.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers and the ACs. Additionally, the authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart, the European Research Council under grant number 683008 (Amplify), and the National Science Foundation (NSF) under grant IIA-1358096.

REFERENCES

1. Anthony D. Andre and Henry A. Cutler. 1998. Displaying Uncertainty in Advanced Navigation Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42, 1 (1998), 31–35. DOI: http://dx.doi.org/10.1177/154193129804200108

2. Peter W. Battaglia, Robert A. Jacobs, and Richard N. Aslin. 2003. Bayesian integration of visual and auditory

signals for spatial localization. *Journal of the Optical Society of America A* 20, 7 (Jul 2003), 1391–1397. DOI: http://dx.doi.org/10.1364/JOSAA.20.001391

3. Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. 2012. A Review of Uncertainty in Data Visualization. *Expanding the Frontiers of Visual Analytics and Visualization* (2012), 81–109. DOI: http://dx.doi.org/10.1007/978-1-4471-2804-5_6
4. David V. Budescu. 2006. Confidence in aggregation of opinions from multiple sources. In *Information Sampling and Adaptive Cognition*. 327–352. DOI: http://dx.doi.org/10.1017/CBO9781107415324.004
5. David V. Budescu, Stephen Broomell, and Han-Hui Por. 2009. Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological science* 20, 3 (2009), 299–308. DOI: http://dx.doi.org/10.1111/j.1467-9280.2009.02284.x
6. Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151. DOI: http://dx.doi.org/10.1109/TVCG.2014.2346298
7. Helen Couclelis. 2003. The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS* 7, 2 (2003), 165–175. DOI: http://dx.doi.org/10.1111/1467-9671.00138
8. Marc O. Ernst. 2010. Decisions Made Better. *Science* 329, 5995 (2010), 1022–1023. DOI: http://dx.doi.org/10.1126/science.1194920
9. Marc O. Ernst and Heinrich H. Bülthoff. 2004. Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8, 4 (2004), 162 – 169. DOI: http://dx.doi.org/10.1016/j.tics.2004.02.002
10. Nivan Ferreira, Danyel Fisher, and Arnd C. Konig. 2014. Sample-oriented Task-driven Visualizations: Allowing Users to Make Better, More Confident Decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. 571–580. DOI: http://dx.doi.org/10.1145/2556288.2557131
11. Nahum Gershon. 1998. Visualization of an imperfect world. *IEEE Computer Graphics and Applications* 18, 4 (1998), 43–45. DOI: http://dx.doi.org/10.1109/38.689662
12. Gerd Gigerenzer, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo, and Konstantinos V. Katsikopoulos. 2005. "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis* 25, 3 (jun 2005), 623–629. DOI: http://dx.doi.org/10.1111/j.1539-6924.2005.00608.x
13. Miriam Greis, Passant El. Agroudy, Hendrik Schuff, Tonja Machulla, and Albrecht Schmidt. 2016. Decision-Making Under Uncertainty: How the Amount of Presented Uncertainty Influences User Behavior. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. 52:1–52:4. DOI:http://dx.doi.org/10.1145/2971485.2971535
14. Miriam Greis, Emre Avci, Albrecht Schmidt, and Tonja Machulla. 2017a. Increasing Users' Confidence in Uncertain Data by Aggregating Data from Multiple Sources. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, 828–840. DOI: http://dx.doi.org/10.1145/3025453.3025998
15. Miriam Greis, Jessica Hullman, Michael Correll, Matthew Kay, and Orit Shaer. 2017b. Designing for Uncertainty in HCI: When Does Uncertainty Help?. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. 593–600. DOI: http://dx.doi.org/10.1145/3027063.3027091
16. Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. 2015. Investigating Representation Alternatives for Communicating Uncertainty to Non-experts. In *Human-Computer Interaction - INTERACT 2015*. Vol. 9299. 256–263. DOI: http://dx.doi.org/10.1007/978-3-319-22723-8_21
17. Miriam Greis, Hendrik Schuff, Marius Kleiner, Niels Henze, and Albrecht Schmidt. 2017. Input Controls for Entering Uncertain Data. *Proceedings of the ACM on Human-Computer Interaction* 1, 1 (jun 2017), 1–17. DOI: http://dx.doi.org/10.1145/3095805
18. Harald Ibrekk and M. Granger Morgan. 1987. Graphical Communication of Uncertain Quantities to Nontechnical People. *Risk Analysis* 7, 4 (1987), 519–529. DOI: http://dx.doi.org/10.1111/j.1539-6924.1987.tb00488.x
19. Robert A. Jacobs. 1999. Optimal integration of texture and motion cues to depth. *Vision Research* 39, 21 (1999), 3621 – 3629. DOI: http://dx.doi.org/10.1016/S0042-6989(99)00088-7
20. Susan L. Joslyn and Jared E. LeClerc. 2012. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology. Applied* 18, 1 (mar 2012), 126–40. DOI: http://dx.doi.org/10.1037/a0025185
21. Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. 2015. Displayed Uncertainty Improves Driving Experience and Behavior: The Case of Range Anxiety in an Electric Car. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*. 2201–2210. DOI: http://dx.doi.org/10.1145/2702123.2702479
22. Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*. 5092–5103. DOI: http://dx.doi.org/10.1145/2858036.2858558

23. Matthew Kay, Dan Morris, Mc Schraefel, and Julie A. Kientz. 2013. There's No Such Thing as Gaining a Pound: Reconsidering the Bathroom Scale User Interface. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. 401–410. DOI: http://dx.doi.org/10.1145/2493432.2493456

24. Isaac M. Lipkus, Greg Samsa, and Barbara K. Rimer. 2001. General performance on a numeracy scale among highly educated samples. *Medical decision making : an international journal of the Society for Medical Decision Making* 21, 1 (2001), 37–44. DOI: http://dx.doi.org/10.1177/0272989X0102100105

25. Alan M. MacEachren, Anthony Robinson, and Susan Hopper. 2005. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* 32, 3 (jan 2005), 139–160. DOI: http://dx.doi.org/10.1559/1523040054738936

26. Rebecca E. Morss, Julie L. Demuth, and Jeffrey K. Lazo. 2008. Communicating Uncertainty in Weather Forecasts: A Survey of the U.S. Public. *Weather and Forecasting* 23, 5 (oct 2008), 974–991. DOI: http://dx.doi.org/10.1175/2008WAF2007088.1

27. Alex T. Pang, Craig M. Wittenbrink, and Suresh K. Lodha. 1997. Approaches to uncertainty visualization. *The Visual Computer* 13, 8 (1997), 370–390. DOI: http://dx.doi.org/10.1007/s003710050111

28. Kristin Potter, Joe Kniss, Richard Riesenfeld, and Chris R. Johnson. 2010. Visualizing Summary Statistics and Uncertainty. *Computer Graphics Forum* 29, 3 (June 2010), 823–832. DOI: http://dx.doi.org/10.1111/j.1467-8659.2009.01677.x

29. Kristin Potter, Paul Rosen, and Chris R. Johnson. 2012. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *IFIP Advances in Information and Communication Technology*, Vol. 377 AICT. 226–247. DOI: http://dx.doi.org/10.1007/978-3-642-32677-6_15

30. Maria H. Ramos, Schalk J. Van Andel, and Florian Pappenberger. 2013. Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences* 17, 6 (June 2013), 2219–2232. DOI: http://dx.doi.org/10.5194/hess-17-2219-2013

31. Gordan Ristovski, Tobias Preusser, Horst K. Hahn, and Lars Linsen. 2014. Uncertainty in medical visualization: Towards a taxonomy. *Computers and Graphics (Pergamon)* 39, 1 (2014), 60–73. DOI: http://dx.doi.org/10.1016/j.cag.2013.10.015

32. Mark S. Roulston, Gary E. Bolton, Andrew N. Kleit, and Addison L. Sears-Collins. 2006. A Laboratory Study of the Benefits of Including Uncertainty Information in Weather Forecasts. *Weather and Forecasting* 21, 1 (2006), 116–122. DOI:http://dx.doi.org/10.1175/WAF887.1

33. Mark S. Roulston and Todd R. Kaplan. 2009. A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts. *Meteorological Applications* 16, 2 (2009), 237–244. DOI: http://dx.doi.org/10.1002/met.113

34. Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 2006. Visualization of Uncertainty in Context Aware Mobile Applications. In *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '06)*. 247–250. DOI: http://dx.doi.org/10.1145/1152215.1152267

35. Orit Shaer, Oded Nov, Lauren Westendorf, and Madeleine Ball. 2017. Communicating Personal Genomic Information to Non-experts: A New Frontier for Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction* 11, 1 (2017), 1–62. DOI: http://dx.doi.org/10.1561/1100000067

36. Meredith Skeels, Bongshin Lee, Greg Smith, and George Robertson. 2008. Revealing uncertainty for information visualization. In *Proceedings of the working conference on Advanced visual interfaces - AVI '08*, Vol. 9. 376. DOI: http://dx.doi.org/10.1145/1385569.1385637

37. Susanne Tak and Alexander Toet. 2014. Color and Uncertainty: It is not always Black and White. *Eurographics Conference on Visualization* (2014), 55–59. DOI:http://dx.doi.org/10.2312/eurovisshort.20141157

38. J. Taylor. 1997. *Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements, 2nd Edition*. University Science Books.

39. Judi Thomson, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. 2005. A typology for visualizing uncertainty. In *IS&T/SPIE Electronic Imaging*, Vol. 5669. 146. DOI: http://dx.doi.org/10.1117/12.587254

40. Amos Tversky and Daniel Kahneman. 1975. Judgment under Uncertainty: Heuristics and Biases. In *Utility, Probability, and Human Decision Making*. Springer Netherlands, Dordrecht, 141–162. DOI: http://dx.doi.org/10.1007/978-94-010-1834-0_8

41. Thomas S. Wallsten, David V. Budescu, Amnon Rapoport, Rami Zwick, and Barbara Forsyth. 1986. Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General* 115, 4 (1986), 348–365. DOI: http://dx.doi.org/10.1037//0096-3445.115.4.348

42. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. 143–146. DOI: http://dx.doi.org/10.1145/1978942.1978963

43. Torre Zuk and Sheelagh Carpendale. 2006. Theoretical analysis of uncertainty visualizations. *Electronic Imaging 2006* 6060, March (jan 2006), 1–14. DOI: http://dx.doi.org/10.1117/12.643631