Detecting Uncertain Input Using Physiological Sensing and Behavioral Measurements

Miriam Greis, Jakob Karolus, Hendrik Schuff, Paweł W. Woźniak, Niels Henze
VIS, University of Stuttgart, Germany
firstname.lastname@vis.uni-stuttgart.de

ABSTRACT
Interactive systems, such as online search interfaces, require appropriate input if they are to produce accurate information. Without this, they can be inaccurate if the user is uncertain about the keywords. Current systems do not have the means to detect uncertainty, which may lead to a negative user experience. We explore physiological and behavioral measurements as tools to implicitly detect users' uncertainty, and provide a method to integrate input variability in interactive systems. We conducted a laboratory study where participants answered questions of varying difficulty, recording physiological data via a key logger, an eye tracker, and a heart rate sensor. Our results show that participants spent significantly more time on difficult questions and looked longer at their answers before submitting them. Based on our results, we provide initial insights on how data from physiological sensors and logged user behavior can be utilized to enrich interactive systems and evaluate a user's uncertainty level.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords
Uncertainty; User Input; Interactive Systems

INTRODUCTION & RELATED WORK
Recent developments in mobile sensing in search technologies empower users to quickly and efficiently obtain information on demand. The majority of the more than one billion[1] iPhone units sold feature a voice-operated personal assistant. Users can easily submit search queries, request directions, or obtain sports results. However, it is not always easy to ask the right questions. Current systems cannot support information retrieval when the user does not communicate their need in a format acceptable by the device. Thus, when one is uncertain of what input to provide, it may be hard to obtain relevant information efficiently. In this paper, we investigate how systems can detect uncertain input using physiological sensing or behavioral measurements. We envision use cases where interactive systems can offer additional support to the user (e.g. predictive input) when the user is hesitant about what command to provide.

Figure 1. The study setup. Participants were asked to answer a set of questions of varying difficulty while seated at a desktop computer.

Although uncertainty is a concept that is present everywhere in daily life, it is seldom taken into account when designing interactive systems. One big strand of research about uncertainty explores and compares visualizations for uncertain data such as glyphs [17] or box plots [20]. Another strand of research deals with the psychological aspects and difficulties of communicating uncertainty to laymen [14, 27]. Recently, research in HCI starts to focus on uncertainty communication to for example support the exploration of genomics data [24] or enhance bus arrival predictions [13]. To be able to communicate and visualize data, the uncertainty has to be quantified, thus also modelled correctly [3]. One of the main challenges for quantifying uncertainty is the user input, which is seldom taken into account. If so, it deals with the technical aspects of uncertainty such as measurement error at the sensor level (e.g. touch screens [22, 28] or capacitive sensing arrays [21]) or the explicit input of uncertainty by the user [8].

Concurrently, the use of ubiquitous physiological sensing is gaining momentum in the HCI field. As sensing devices become parts of everyday life, e.g. as parts of smart watches,

[1] http://www.apple.com/newsroom/2016/07/apple-celebrates-one-billion-iphones.html

MUM 2017, November 26–29, 2017, Stuttgart, Germany

ACM ISBN 978-1-4503-5378-6/17/11…$15.00

DOI: https://doi.org/10.1145/3152832.3152859

research is looking for new use cases for extended physiological data about the user. While heart rate has been extensively used by sportsmen to understand the details of their training performance [25], Curmi et al. [5] propose using heart rate to enhance social interactions. Recently, McDuff et al. [16] demonstrated that adaptive systems can infer the level of cognitive stress based on the user's heart rate variability and adjust interaction properties accordingly. In a video conference scenario, Hosoda et al. [9] used heart rate resonance measurements to assess the level of engagement of participants. In our work, we investigate if heart rate data can be used to measure how uncertain the user is about the input they are providing.

Further, we also look at additional information that can be extracted from the user's gaze pattern. Vision is one of our most important senses. We predominantly focus our attention and gaze direction on elements of interest. Researchers have used this fact to develop applications for gaze-contingent interfaces, such as the creation of photo albums by rating pictures on how often they were looked at [26] or tutoring systems [6] and translation assistance [12]. Furthermore, gaze patterns have been analyzed in quiz [4] and problem solving studies [15] indicating varying patterns when users are confronted with unfamiliar or difficult tasks.

In this paper, we present a study where we asked 21 participants to answer questions of varying difficulty on a desktop computer while we measured their heart rate and eye gaze (see Figure 1). In order to do so, we compiled a set of general knowledge questions with adjusted difficulty. We then identified a subset of measures that produced significant effects and may be used for uncertainty detection. Thus, we contribute initial insights for building uncertainty-aware systems with physiological sensing. Next, we discuss the implications of our study for future systems.

This paper contributes the following:

1. A study of how the participants' physiological and behavioural parameters vary when answering questions of varying difficulty when seated in front of a desktop computer;
2. A rich data set of physiological measurements. The dataset is publicly available and intended for reuse in future research;

METHOD

We conducted a study in order to evaluate the influence of users' uncertainty on physiological signals and behavioral measurements. Participants had to answer 140 questions with different difficulty levels and subsequently self-reported their perceived uncertainty. We chose an ex-post-facto experimental design as the self-reported score cannot be controlled directly, but has to be determined after the participants answered the questions.

Question Selection Process

To identify questions with different levels of uncertainty, we built a pool of free-text questions following a three-step selection process. We first transcribed 1770 German quiz questions from four books about general knowledge [2, 10, 11, 18]. As they were multiple choice questions, we manually sorted them and deleted questions that were not solvable without multiple choice answers. On the remaining 1164 free-text questions, we applied the following filter criteria:

- **Maximum answer length of four words:** We removed all questions containing answers of more than four words, eliminating full-sentence answers. Thus, we minimize the confounding uncertainty resulting from the need to spell long complex phrases.
- **Maximum question length of 15 words:** All questions with more than 15 words were removed as this is the upper border for the recommended sentence length in German for easy comprehension [23]. Thus, we remove complex sentences that could be difficult to read.
- **Flesch-Reading-Ease of questions between 60 and 80:** The Flesch-Reading-Ease [7] (FRE) is a readability metric that measures how difficult it is to understand a text based on average sentence-length (ASL) and average number of syllables per word (ASW). We use an adapted version ($FRE_{german} = 180 - ASL - (58.5 \cdot ASW)$) for German language as proposed by Amstadt [1] to filter very easy or very difficult questions and reduce the confounding uncertainty introduced by complex sentences. Sentences with a FRE of 60 to 80 fall into the category of medium and medium easy sentences.

We conducted an online survey including the remaining 251 questions. In total, 59 participants provided 7939 answers to questions ($M = 134.55$ questions per participants, $SD = 102.7$). For each question, participants had to specify how certain their answer was on a 5-point Likert scale. We assigned each question an uncertainty class corresponding to the item most participants selected on the Likert scale. For each item on the Likert scale, we picked 40 questions for the extremes and respectively 20 questions for the items in the middle. We picked the questions for the class by calculating the ratio of number of assignments to the class and the total answers for the question taking questions with a higher ratio.

Participants

We recruited 24 participants (15 male, 9 female) with an average age of 23.17 ($SD = 3.36$). All of them were native German speakers. For our analysis, we used the data of 20 participants as some data was lost due to technical difficulties. For the analysis of the eye movements and the heart rate, we used subsets of the participant base due to unreliable tracking caused by make-up and loosened electrodes.

Apparatus

The study was conducted using a remote eye-tracker (SMI RED 250) attached to the bottom of a 22 in. LCD-display. To enter their answers and evaluations, participants were given a keyboard and a mouse. In addition, we used three dividing walls to shield the participants from disturbances. To record the ECG-signal, we used a NEXUS connected to the recording laptop via Bluetooth. The interface presenting the questions was implemented as a website running in the browser.

Procedure

After arriving at the lab, participants filled a consent form and a demographic questionnaire. The three ECG-electrodes were attached to the left and right arm and above the left ankle. After calibrating the eye tracker, participants got detailed instructions on the tasks. We asked participants to provide a reasonable answer (e.g., stating a city name if the question asked for a city) even if they did not know the correct answer. After pressing start and answering a test question, participants were shown all 140 questions in randomized order (see Figure 2 for a sample question) answering them one by one.

Figure 2. Example question presented to the participants. Translation: How were craftsmen called in Athen?

Throughout the study, we collected data from the eye-tracker, the NEXUS, and the browser. Eye movements were recorded at 250 Hz, the ECG-signal at 256 Hz. Regarding the browser data, we collected all key events, click events, mouse movements, answer completion times, the participants' answers, and field focus events.

After answering all questions, participants rated their uncertainty levels for each presented questions based on a 5-point Likert scale from "Entirely don't agree" to "Fully agree" given the statement: "I am sure that my answer is correct." (see Figure 3). The order of questions was again randomized.

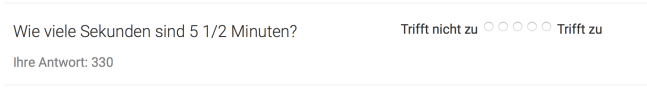

Figure 3. Exemplary evaluation for one question. Translation: How many seconds are 5 1/2 minutes? Your answer: 330

Measures

Using the collected data we derived several metrics from each data source. Regarding the ECG-signal, we calculated heart rate and heart rate variability. From the browser data we extracted numerous features concerning time (such as completion time, time between first and last typing), typing behaviour (such as typing speed, key down time, and deletion count) and mouse events (such as the length of the mouse path and click counts). From our recorded eye movement data, we extracted features related to fixations, saccades, gaze direction and eye blinks. For fixations, saccades, and blinks we mainly looked at duration (average, total), count as well as velocity metrics (e.g. acceleration for saccades). We coarsely analyzed gaze direction by measuring the amount of time that the user spent on specific screen elements, such as the question itself and the answer field. Additionally, we submitted measures related to refixation ratio and backpropagations for statistical analysis.

RESULTS

In this section, we present statistical results for a subset of the measures in our study. We focus on data that we identified as highly promising for detecting uncertainty. We provide examples of data from the data sources.

Browser data

Time between first and last typing

First, we look at the total time elapsed from when the user began typing the answer to when the user finished typing. The grand mean was 4.43 s (SD = 7.13 s). Questions with the lowest self-perceived uncertainty (very certain) required the least time (M = 3.73 s, SD = 5.60 s) while the one with the highest self-perceived uncertainty (very uncertain) required the longest typing periods (M = 5.51 s, SD = 9.91 s). We conducted a one-way ANOVA to investigate the effect of question uncertainty level on time spent between first and last typing. The effect was statistically significant ($F_{4,2717}$ = 8.34, $p < .001$).

Time before first typing

We also investigated the time elapsed before the users started typing their answers (see Figure 4). We recorded the grand mean at 9.19 s (SD = 9.45 s). For self-perceived lowest uncertainty, participants required the lowest amount of time to begin providing an answer (M = 5.15 s, SD = 4.56 s). The highest self-perceived uncertainty questions produced the longest times before typing (M = 13.14 s, SD = 13.07 s). We conducted a one-way ANOVA to investigate the effect of question uncertainty level on time elapsed before the users began typing. The effect was statistically significant with $F_{4,2717}$ = 84.01 and $p < .001$.

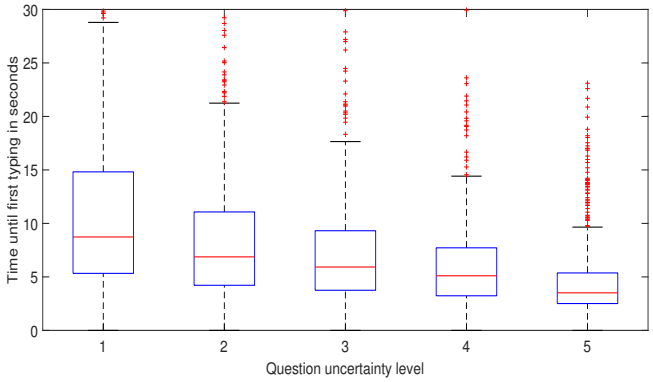

Figure 4. Boxplot of time elapsed until first typing grouped by questions uncertainty level (1: very uncertain, 5: very certain).

Eyetracker data

Time spent looking at the answer field

Next, we present the analysis for the time users spent looking at the answer field, normalized by the total time looking at the screen. The average time ratio was 0.30 (SD = 0.19). Participants spent the most amount of time looking at answer field when solving "very certain" questions (M = 0.34,

SD = 0.21). The "very uncertain" questions caused participants to look at the answer field for the shortest time (M = 0.26, SD = 0.18). We conducted a one-way repeated ANOVA to investigate the effect of question uncertainty level on time spent between looking at the answer field. We observed a statistically significant effect ($F_{4,1907}$ = 13.53, $p < .001$).

Refixation ratio

Lastly, we investigate the amount of refixations that occurred, normalized by the fixation count for the respective question (see Figure 5). The average refixation ratio was 0.36 (SD = 0.17). The ratio was lowest for the "very certain" questions (M = 0.30, SD = 0.17). The "very uncertain" questions showed the highest ratio (M = 0.41, SD = 0.18). We conducted a one-way repeated ANOVA to investigate the effect of question uncertainty level on the refixation ratio. We observed a statistically significant effect ($F_{4,1907}$ = 32.03, $p < .001$).

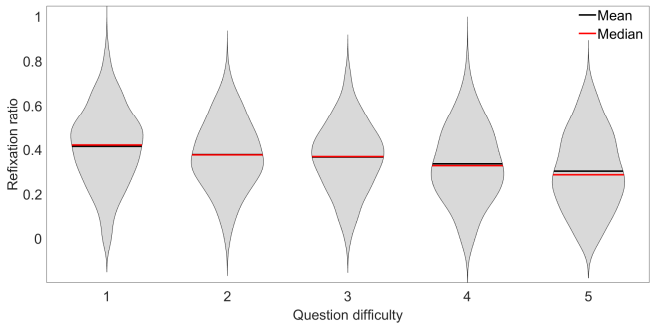

Figure 5. Violin plot of refixation ratio grouped by questions difficult (1: very uncertain, 5: very certain).

ECG data

We investigated median heart rates (HR) and heart rate variability (HRV) while answering questions. As it is understood that heart rate reacts with a certain delay, we tried multiple aggregation strategies to determine if uncertainty had an effect on heart behaviour. We used combinations of four different lag values — $1000, 5000, 9000, 0\,[ms]$ and three different summation window sizes — $1000, 5000, 10000\,[ms]$, resulting in 12 sets of measurements. We performed one-way ANOVAs to investigate the effect of question difficulty on median HR and HRV. No significant results were found for both measures ($p > .05$)

DISCUSSION

The results of our study indicate that there are multiple possibilities of sensing uncertainty of user input. While we focused on physiological sensing, we found significant differences in measures obtained using the keyboard. It is possible that extensive computational analysis of keyboard behaviour may be enough to measure uncertainty in some scenarios. However, it appears that combining data from eye tracking and keyboard input can provide a more reliable metric. This question, however, is outside of the scope of this paper. We make our full data set available to the research community to explore this question further.

We were surprised to learn that question difficulty did not have a significant effect on heart rate measurements or related eye tracking metrics such as time spent looking at the screen. As past work indicated that heart rate was correlated with cognitive stress [16] and conversational engagement [9], we expected that hesitation would produce a similar effect. We believe that our results can be explained by two reasons. Firstly, our study offered a comfortable and safe environment for the users. As the participants knew that their answers did not have any implications, being unsure of an answer did not produce detectable physiological effects including averting one's gaze[2] during a thought process. A way to address this issue in a future study would be to reward participants based on performance. Secondly, heart rate reactions to hesitation may be delayed by a time period that is longer than the answer to the question we provided. This appears to be likely as it takes 5 or more seconds for the heart rate to react to sudden physical exertion [19].

We recognise that our study is prone to certain limitations. We used only questions in German, which may have created a cultural bias. It is possible that other cultures react to hesitation in a different way, producing stronger physiological signals. Our experiment was performed in a controlled environment where outside distractions were minimised. We do not know if external factors could affect the physiological response to uncertainty. Furthermore, we used only three sensing modalities in our work. Future research should explore if other sensors could provide better results.

CONCLUSION

In this work, we presented our inquiry into detecting uncertainty in input through physiological sensing and user behavior data. We reported the results of our experiment where participants answered questions of varying difficulty. We gathered data from a browser, an eye tracker and a heart rate sensor and analysed it to identify measures useful for approximating uncertainty. Time between first and last typing, and time spent looking at the question were identified as useful metrics that can be obtained from a browser. We also pointed out that the time of looking at the answer field and refixation count are relevant measures that can be acquired with an eye tracker. Finally, we found that heart rate was unlikely to provide information useful in determining if the user is hesitant. Thus, we contribute a first step in determining which data sources may be useful for building uncertainty-aware systems.

To support further research, we will release the physiological measurement data obtained to the public domain. In future work, we plan to use artificial intelligence tools and machine learning to build advanced methods of uncertainty detection. We hope that the work presented in this paper will inspire further research in providing additional, contextualised input assistance when users are uncertain about how to interact with a system.

[2]Time spent looking at the screen did not significantly differ for the different uncertainty levels.

ACKNOWLEDGMENTS

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.

REFERENCES

1. Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service.
2. Antonia Bauer and Ansbert Kneip. 2016. *Der große Wissenstest für Kinder - Was weißt du über die Welt?*
3. Nadia Boukhelifa and David John Duke. 2009. Uncertainty Visualization: Why Might It Fail?. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. ACM, New York, NY, USA, 4051–4056. DOI: **http://dx.doi.org/10.1145/1520340.1520616**
4. Leana Copeland and Tom Gedeon. 2013. The Effect of Subject Familiarity on Comprehension and Eye Movements During Reading. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration (OzCHI '13)*. ACM, New York, NY, USA, 285–288. DOI:**http://dx.doi.org/10.1145/2541016.2541082**
5. Franco Curmi, Maria Angela Ferrario, Jen Southern, and Jon Whittle. 2013. HeartLink: Open Broadcast of Live Biometric Data to Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1749–1758. DOI: **http://dx.doi.org/10.1145/2470654.2466231**
6. Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System. *International Journal of Human-Computer Studies* 70, 5 (May 2012), 377–398. DOI: **http://dx.doi.org/10.1016/j.ijhcs.2012.01.004**
7. Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221. DOI: **http://dx.doi.org/10.1037/h0057532**
8. Miriam Greis, Hendrik Schuff, Marius Kleiner, Niels Henze, and Albrecht Schmidt. 2017. Input Controls for Entering Uncertain Data: Probability Distribution Sliders. *Proc. ACM Hum.-Comput. Interact.* 1, 1, Article 3 (June 2017), 17 pages. DOI: **http://dx.doi.org/10.1145/3095805**
9. Masamichi Hosoda, Akira Nakayama, Minoru Kobayashi, and Satoshi Iwaki. 2004. Conference State Estimation by Biosignal Processing: Observation of Heart Rate Resonance. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. ACM, New York, NY, USA, 1187–1190. DOI: **http://dx.doi.org/10.1145/985921.986020**
10. Jürgen Hotz. 2013. *Duden - Testen Sie Ihre Allgemeinbildung*.
11. Jürgen Hotz. 2014. *Duden - Testen Sie Ihre Allgemeinbildung 2*.
12. Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. 2003. Proactive Response to Eye Movements. In *INTERACT*, Vol. 3. IOS press Amsterdam, 129–136.
13. 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 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5092–5103. DOI: **http://dx.doi.org/10.1145/2858036.2858558**
14. Isaac M. Lipkus, Greg Samsa, and Barbara K. Rimer. 2001. General Performance on a Numeracy Scale among Highly Educated Samples. *Medical Decision Making* 21, 1 (2001), 37–44. DOI: **http://dx.doi.org/10.1177/0272989X0102100105**
15. Adrian Madsen, Adam Larson, Lester Loschky, and N. Sanjay Rebello. 2012. Using ScanMatch Scores to Understand Differences in Eye Movements Between Correct and Incorrect Solvers on Physics Problems. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. ACM, New York, NY, USA, 193–196. DOI: **http://dx.doi.org/10.1145/2168556.2168591**
16. Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 2016. COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4000–4004. DOI: **http://dx.doi.org/10.1145/2858036.2858247**
17. 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**
18. Heike Pfersdorff and Iris Glahn. 2015. *Duden - Testen Sie Ihr Wissen! Das Allgemeinbildungsquiz.*
19. Pete Pfitzinger and Scott Douglas. 2001. *Advanced marathoning*. Human Kinetics.
20. Kristin Potter, Joe Kniss, Richard Riesenfeld, and Chris R. Johnson. 2010. Visualizing Summary Statistics and Uncertainty. *Computer Graphics Forum* 29, 3 (2010), 823–832. DOI:**http://dx.doi.org/10.1111/j.1467-8659.2009.01677.x**
21. Simon Rogers, John Williamson, Craig Stewart, and Roderick Murray-Smith. 2010. FingerCloud: Uncertainty and Autonomy Handover Incapacitive Sensing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 577–580. DOI: **http://dx.doi.org/10.1145/1753326.1753412**

22. Julia Schwarz, Scott Hudson, Jennifer Mankoff, and Andrew D. Wilson. 2010. A Framework for Robust and Flexible Handling of Inputs with Uncertainty. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 47–56. DOI: **http://dx.doi.org/10.1145/1866029.1866039**

23. Wilfried Seibicke. 1969. *Duden - Wie schreibt man gutes Deutsch? Eine Stilfibel.*

24. Orit Shaer, Oded Nov, Johanna Okerlund, Martina Balestra, Elizabeth Stowell, Lauren Westendorf, Christina Pollalis, Jasmine Davis, Liliana Westort, and Madeleine Ball. 2016. GenomiX: A Novel Interaction Tool for Self-Exploration of Personal Genomic Data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 661–672. DOI: **http://dx.doi.org/10.1145/2858036.2858397**

25. Jakob Tholander and Stina Nylander. 2015. Snot, Sweat, Pain, Mud, and Snow: Performance and Experience in the Use of Sports Watches. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2913–2922. DOI: **http://dx.doi.org/10.1145/2702123.2702482**

26. Tina Caroline Walber, Ansgar Scherp, and Steffen Staab. 2014. Smart Photo Selection: Interpret Gaze As Personal Interest. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2065–2074. DOI: **http://dx.doi.org/10.1145/2556288.2557025**

27. 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**

28. Daryl Weir. 2012. Machine Learning Models for Uncertain Interaction. In *Adjunct Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct Proceedings '12)*. ACM, New York, NY, USA, 31–34. DOI: **http://dx.doi.org/10.1145/2380296.2380313**