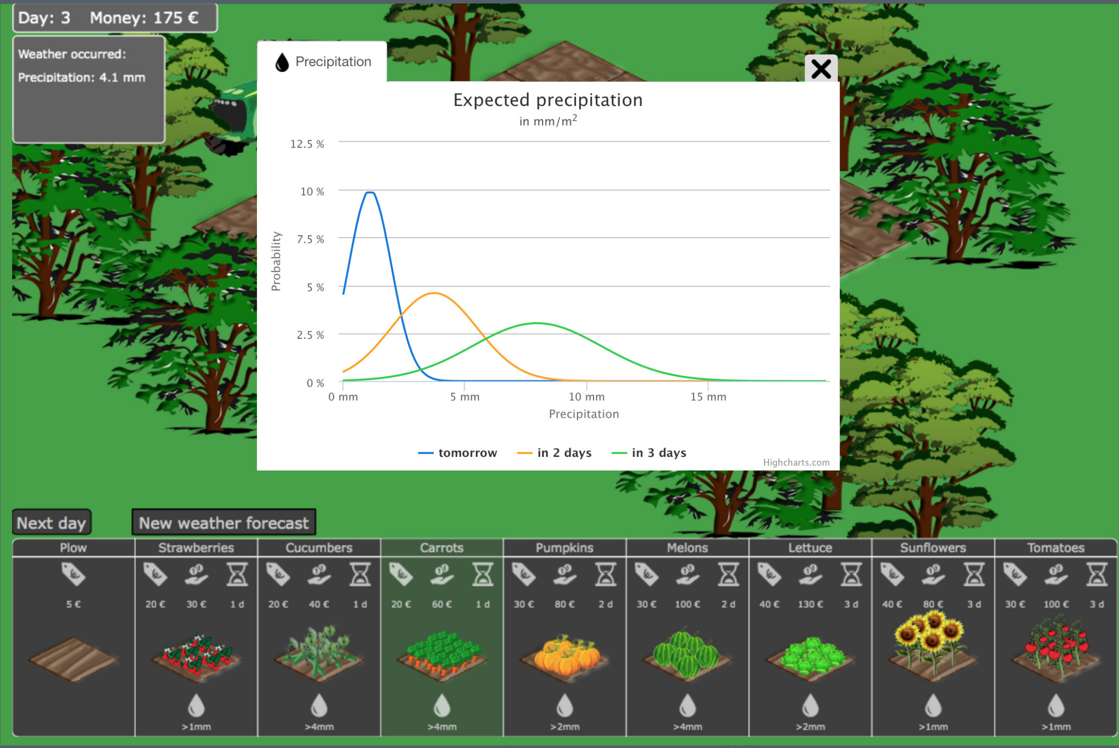
A Systematic Exploration of Uncertainty in Interactive Systems

Miriam Greis



Day: 3 Money: 175 €

Weather occurred:
Precipitation: 4.1 mm

Plow	Strawberries	Cucumbers	Carrots	Pumpkins	Melons	Lettuce	Sunflowers	Tomatoes
5 €	20 €	20 €	20 €	30 €	30 €	40 €	40 €	30 €
	30 €	40 €	60 €	80 €	100 €	130 €	80 €	100 €
	1 d	1 d	1 d	2 d	2 d	3 d	3 d	3 d
	>1mm	>4mm	>4mm	>2mm	>4mm	>2mm	>1mm	>1mm

University of Stuttgart
Germany

A SYSTEMATIC EXPLORATION OF UNCERTAINTY IN INTERACTIVE SYSTEMS

Von der Fakultät für Informatik, Elektrotechnik und Informationstechnik und dem Stuttgart Research Centre for Simulation Technology (SRC SimTech) der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

vorgelegt von

MIRIAM GREIS

aus Esslingen am Neckar

Hauptberichter:	Prof. Dr. Albrecht Schmidt
Mitberichter:	Prof. Dr. Orit Shaer
Tag der mündlichen Prüfung:	21. Dezember 2017

Institut für Visualisierung und Interaktive Systeme
der Universität Stuttgart

2017

Abstract

Uncertainty is an inherent part of our everyday life. Humans have to deal with uncertainty every time they make a decision. The importance of uncertainty additionally increases in the digital world. Machine learning and predictive algorithms introduce statistical uncertainty to digital information. In addition, the rising number of sensors in our surroundings increases the amount of statistically uncertain data, as sensor data is prone to measurement errors. Studies in Psychology have revealed that humans prefer to receive information about uncertainty and make better decisions if uncertainty information is communicated. Furthermore, an adequate communication of uncertainty establishes trust in an application. Hence, there is an emergent need for practitioners and researchers in Human-Computer Interaction to explore new concepts and develop interactive systems able to handle uncertainty. Such systems should not only support users in entering uncertainty in their input, but additionally present uncertainty in a comprehensible way.

The main contribution of this thesis is the exploration of the role of uncertainty in interactive systems and how novel input and output methods can support researchers and designers to efficiently and clearly communicate uncertainty. By using empirical methods of Human-Computer Interaction and a systematic approach, we present novel input and output methods that support the comprehensive communication of uncertainty in interactive systems. We further integrate our results in a simulation tool for end-users.

First, we identify functional and non-functional requirements for the interaction with uncertainty in the context of an end-user simulation tool. We conduct multiple surveys, focus groups, and design workshops with simulation experts and end-users. The results of our studies show that delivering designs that cater to different users and application scenarios, and transparency of used algorithms and data, play an important role. Both can only be achieved by an adequate quantification and communication of uncertainty.

Based on related work, we create a systematic overview of sources of uncertainty in interactive systems to support the quantification of uncertainty and identify relevant research areas. The overview can help practitioners and researchers to identify uncertainty in interactive systems and either reduce or communicate it. We then introduce new concepts for the input of uncertain data. We enhance standard input controls, develop specific slider controls and tangible input controls, and collect physiological measurements. For each of these explorations, we conduct surveys and studies to identify and propose the most promising candidates for future usage. We also compare different representations for the output of uncertainty to make recommendations for their usage. Furthermore, we analyze how humans interpret uncertain data und make suggestions on how to avoid misinterpretation and statistically wrong judgements.

We embed the insights gained from the results of this thesis in an end-user simulation tool to make it available for future research. The tool is intended to be a starting point for future research on uncertainty in interactive systems and foster communicating uncertainty and building trust in the system.

Overall, our work shows that user interfaces can be enhanced to effectively support users with the input and output of statistically uncertain information.

ZUSAMMENFASSUNG

Unsicherheit war und ist ein inhärenter Teil des Alltags und Menschen werden damit täglich konfrontiert, wenn sie Entscheidungen treffen. Zunehmend spielt das Thema Unsicherheit auch eine immer größere Rolle in der digitalen Welt. Durch die Entwicklung von Algorithmen, die maschinell lernen und Vorhersagen treffen, sind viele digitale Informationen statistisch mit Unsicherheit behaftet. Auch die wachsende Anzahl an Sensoren in der Umgebung führt zu einem Zuwachs an statistisch unsicheren Daten, da die Sensordaten anfällig für Messungenauigkeiten sind. Studien in der Psychologie haben bereits gezeigt, dass Menschen es bevorzugen, über Unsicherheit informiert zu werden, und besser Entscheidungen treffen, wenn Informationen über die Unsicherheit zur Verfügung stehen. Auch das Vertrauen in Informationen kann durch eine adäquate Kommunikation der Unsicherheit gesteigert werden. Gerade deshalb ist es umso wichtiger, dass Entwickler und Forscher in der Mensch-Computer-Interaktion Konzepte erforschen und interaktive Systeme entwickeln, die mit Unsicherheit umgehen können. Diese Systeme sollen dem Benutzer einerseits erlauben Unsicherheit bei der Eingabe zu kommunizieren und andererseits Unsicherheit auf verständliche Weise darstellen.

Der Hauptbeitrag dieser Arbeit liegt in der Erforschung, inwiefern Unsicherheit in interaktiven Systemen eine Rolle spielt und wie neue Eingabe- und Ausgabemethoden dazu beitragen können, Unsicherheit effizient und verständlich zu kommunizieren. Mit Hilfe des Einsatzes von empirischen Methoden der Mensch-Computer-Interaktion und einer konstruktiven Vorgehensweise stellen wir neuartige Eingabe- und Ausgabemethoden vor, die die verständliche Kommunikation von Unsicherheit in interaktiven Systemen unterstützen. Unsere Erkenntnisse integrieren wir in einem Simulationswerkzeug für Endnutzer.

Zunächst identifizieren wir notwendige funktionale und nicht-funktionale Voraussetzungen für die Interaktion mit Unsicherheiten im Kontext eines Simulationswerkzeugs, welches für Endnutzer geeignet ist. Dafür führten wir mehrere Umfragen, Fokusgruppen und Designworkshops mit Simulationsexperten und Endnutzern durch. Die Ergebnisse unserer Untersuchungen zeigen, dass Flexibilität in Bezug auf unterschiedliche Benutzer und Anwendungsfälle sowie Transparenz von verwendeten Algorithmen eine große Rolle spielen. Beides kann jedoch nur über die adäquate Quantifizierung und Kommunikation von Unsicherheit erreicht werden.

Um die Quantifizierung von Unsicherheit zu unterstützen und relevante Forschungsbereiche zu identifizieren, stellen wir basierend auf verwandten Arbeiten eine systematische Übersicht über Quellen von Unsicherheit in interaktiven Systemen vor. Diese Übersicht kann zukünftig Forschern und Entwicklern helfen, Unsicherheit in interaktiven Systemen zu identifizieren und entweder zu reduzieren oder zu kommunizieren. Darauf aufbauend präsentieren wir neue Konzepte für die Eingabe von unsicheren Daten. Wir erweiterten dazu Standardeingabemethoden, entwickelten spezifische Slider, begreifbare Eingabemethoden und erhoben physiologische Messungen. Zu jedem Teilbereich führten wir Umfragen und Studien durch, um die Nutzung der vielversprechendsten Eingabemethoden zu empfehlen. Auch für die Ausgabe von unsicheren Daten vergleichen wir verschiedene Repräsentationen, um Empfehlungen zu deren Verwendung auszusprechen. Des Weiteren untersuchen wir, wie Menschen unsichere Daten interpretieren, und sprechen Empfehlungen zur Verhinderung von Missinterpretationen bzw. statistisch falschen Interpretationen aus.

Die Ergebnisse dieser Arbeit werden im Rahmen eines Simulationswerkzeuges für Endnutzer aufgegriffen und für zukünftige Forschung aufbereitet. Dieses Werkzeug soll als Basis für zukünftige Forschung über Unsicherheit in interaktiven Systemen dienen, um die verständliche Kommunikation von Unsicherheit und das Vertrauen in interaktive Systeme zu fördern.

Insgesamt zeigt diese Arbeit, dass Benutzungsschnittstellen sinnvoll so erweitert werden können, dass Benutzer bei der Ein- und Ausgabe von statistisch unsicheren Informationen effektiv unterstützt werden können.

Preface

This thesis originated from the research that I conducted at the University of Stuttgart in the context of a project funded via the Excellence Initiative of the German Research Foundation. My work and decisions were influenced by many conversations and discussions with colleagues, students, and external researchers working on the topic of uncertainty in interactive systems. As a research associate at the University of Stuttgart, I also supervised student projects including Bachelors and Masters theses. These theses were all related to my research topic and supported me in realizing my ideas. During my whole time as PhD student, I very much valued the scientific exchange with other researchers and practitioners when attending conferences, workshops, or doctoral colloquiums. Hence, I decided to write this thesis using the scientific plural instead of the singular. All figures and diagrams in this paper were either made by myself or originated in the context of theses completed under my supervision. Additionally, parts of the presented work are based on scientific publications arising from collaborations with colleagues and students. The respective chapters contain references to these publications in the introductory part of the chapter.

Acknowledgments

While my name is the only one on the front cover of this thesis, there are many others that supported me throughout my time as a PhD student. I had the unique opportunity to meet excellent researchers and collaborators without whom I would not have been able to finish this thesis. I want to thank all of them for influencing and shaping my research. Apart from that, many did not only stay collaborators but turned into great friends who gave me the personal support I needed to keep up and continue my research with all its ups and downs. I therefore dedicate the acknowledgements to them and sincerely apologize to those, that I might have missed.

First and foremost, I would like to express my special appreciation to my supervisor **Albrecht Schmidt** who had faith in employing me as a research assistant offering me the chance to pursue my PhD in his research group. He inspired my work and always supported me in the best possible way to achieve my goals. Thank you, Albrecht, for your continuos support and guidance, your faith in my research despite the many paper rejections I faced during my first two years, and your precious time whenever I needed it most. I could not have imagined a better supervisor for my PhD thesis. Besides my supervisor, I would like to thank my thesis committee **Orit Shaer**, **Thomas Ertl**, and **Stefan Funke** for their insightful comments and questions. Thank you for your feedback, time, and effort!

I started working as a teaching assistant in Albrecht's research group when I was a master student. I am very grateful that he offered me to write a diploma thesis in his group. I would like to thank **Niels Henze** and **Florian Alt**, who supervised my diploma thesis for making me interested in research. Without them, I would have probably not joined Albrecht's group to pursue a PhD.

At the University of Stuttgart, I had many excellent colleagues who supported my research and time as a PhD student in different ways. Most thanks goes to **Paweł W. Woźniak** for his constant personal support since he joined Albrecht's group. Whenever I struggled or needed someone to talk, I knew that I could call him. He mentally supported me in finishing some of the most important milestones for this thesis and always found the most friendliest and motivating words to encourage me to keep up my research. I found a further great collaborator in **Tonja Machulla**, who supervised two students with me, which resulted in an Honorable Mention Award on CHI'17. Special thanks also to **Jakob Karolus**, not only for sharing an office with me during the last 1 1/2 years, but also for numerous cooking and music sessions, which I will truly miss now that he moved to Munich. Thanks goes to **Yomna Abdelrahman** (for Egyptian sweets and awesome food), **Nora Broy**, **Mariam Hassib**, **Romina Kettner**, and **Alexandra Voit** for being the greatest roommates on conferences and seminars, that I could possibly have. It was always a great pleasure to share a hotel room and a conference experience with you. Thanks as well to all other colleagues that I met during my time at the University of Stuttgart: **Mauro Avila**, **Patrick Bader** (for randomly picking the same marriage date), **Céline Coutrix** (for the collaboration on tangible interfaces), **Tilman Dingler** (for great music on our Christmas parties), **Passant El.Agroudy** (for her endless energy), **Markus Funk** (for having trust in me as a skiing teacher), **Huy Viet Le** (for our adventurous visit to the zoo), **Hyunyoung Kim**, **Francisco Kiss**, **Pascal Knierim**, **Oliver Korn** (for realizing that there are women in computer science), **Thomas Kosch**, **Thomas Kubitza**, **Lars Lischke** (for being a great travel buddy on my first CHI conference), **Sven Mayer** (for always fixing the repository of our students and great barbecues), **Bastian Pfleging** (for his invaluable knowledge about everything and insights on how to write a PhD thesis), **Rufat Rzayev** (for always being in the mood for singing a song), **Stefan Schneegaß**, **Valentin Schwind**, **Dominik Weber**, and **Katrin Wolf**. Special thanks goes to **Anja Mebus** and **Murielle Naud-Barthelmeß** for doing all the administrative work and always being receptive for problems arising in the group.

Being a member of the graduate school of the Cluster of Excellence in Simulation Technology introduced me to many people working in different fields. First, I want to say thank you to my project network: **project network 7**. Thank you for great interdisciplinary conversations in our lunch meetings and on the status seminars. Second, my thanks goes to **Maria Hammer** and **Christoph Grüninger** for sharing the honor of being PhD spokespersons during the first year of my PhD and organizing a great PhD weekend! I further want to thank **Mark Dornbach**, **Dennis Grunert**, and **Andreas Schmidt** for our informal lunch meetings on Thursdays. What started as a meeting with PhD students from other faculties lead to a great exchange of experiences and a friendship that I do not want to miss anymore. I additionally want to thank the whole **SimTech management team**, especially **Barbara Teutsch** who did a great job in managing the graduate school. Thank you, Barbara, for always having an open door no matter what questions or problems I had.

There are many other great people that I am grateful for getting to know during my PhD. First of all **Chris Schmandt** and his awesome group at the MIT Media Lab, who made it possible for me to visit them for three months. I enjoyed my stay very much experiencing a different culture of research. Thank you, Chris, for being such a supportive supervisor sharing your great knowledge and stories about the beginnings of the Media Lab. On the numerous conferences I visited, I met many other great and inspiring researchers. I want to thank: **Jonathan Day** (for being the best SV Co-Chair that I could have ever picked) and all our great SVs of UbiComp'16, **Pascal Lessel** (for the shared experience of four CHI conferences and for always being an awesome seatmate on the return flights), **Marion Kölle** (for being my roommate at the winter school and on NordiCHI'16 and for the small surprise package after my defense), **Thijs Roumen** (for the great organization and success of the German party on CHI'17), **Christian Löw** (for being a great motivator on MobileHCI'17), **Michael Lahnert** (for sharing some of the painful experiences of being a PhD student), **Matthew Kay** (for supporting my idea of a CHI workshop on uncertainty), all organizers and participants who made the uncertainty workshop happen on CHI'17 and all the others whose names do not fit in the limited space of these acknowledgements.

This thesis would have not been possible without the support of great students. Thanks to my two student research assistants **Hendrik Schuff** and **Ken Singer** for their great and reliable work over multiple years. I additionally supervised multiple bachelor and master theses to shape my research. Thanks to **Emre Avci**, **Velihan Bulut**, **Patrick Franczak**, **Marius Kleiner**, **Andreas Korge**, and **Thorsten Ohler** for being great students that I really liked to work with. I also got the chance to supervise two exchange students, **Vicky Ziemer** and **Aditi Joshi**, who did an amazing job during the three months they joined us for doing research. It was a pleasure to work with you! Thank you, Vicky, for also supporting me in finding a room in Boston and being a great friend when I visited the MIT Media Lab. I am also very thankful for the great conversations with **Johannes Knittel** and **Dominique Rau** who I already supervised as a teaching assistant during my studies and later met in their role of startup founders.

I also want to thank some of my friends that I met during my time at the University mainly already as a student. Nevertheless, I was always able to discuss questions regarding my PhD thesis with them to get a different perspective and heavily relied on their personal support. Thanks to **Pascal Hirmer**, **Niklas Kaulitz**, **Severin Leonhardt**, **Sebastian Richter**, and **Hansjörg Schmauder** for great lunch breaks, walks, and visits.

Last but not least, I want to thank my family for always supporting me in any possible way to complete this thesis: my parents **Marion Greis** and **Martin Greis**, my sister **Madlen Greis** and especially my husband **Florian Greis**. Without your support, I would have not been the person I am today and I would have not finished this thesis.

Thank you!

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

List of Acronyms

AMS American Meteorological Society

ASL average sentence-length

ASW average syllables per word

CSUQ Computer System Usability Questionnaire

DoF degrees of freedom

FEM finite element method

FRE Flesch-Reading-Ease

FVM finite volume method

GIS geographic information systems

GIScience Geographic Information Science

GNSS global navigation satellite system

GPS Global Positioning System

HCI Human-Computer Interaction

HEPS Hydrological Ensemble Prediction Systems

IC input control

NIST National Institute of Standards and Technology

NRC National Research Council

RQ research question

SUS System Usability Scale

UMUX Usability Metric for User Experience

I

Introduction and Motivation

Chapter 1

Introduction

Nowadays, interactive systems often rely on sensor data or complex mathematical methods such as simulation and machine learning. This development poses new challenges for the developers and designers of such systems. This is caused by the fact that these mathematical methods are difficult to understand for non-experts such as non-scientists. In recent years, research on user-centered machine learning and simulations has started to emerge. However, these movements mostly focus on how to make these methods easier to use for experts and not on how to make it understandable for the general public.

One of the main challenges connected with simulations and machine learning is that these methods produce uncertain data. Uncertainty quantification and visualization is therefore a research field in many different natural sciences. The main focus of this research is often the scientists themselves. Many different fields work on different perspectives of the topic yet there is no consistent definition of the term *uncertainty*, so varying terminologies and taxonomies concentrating on different aspects of the topic are used. In Human-Computer Interaction (HCI), however, the topic of uncertainty is fairly new and has only recently gained the attention of the research community. Uncertainty is an inherent part of the world and something users know from their everyday lives, for example by using weather forecasts. The expectation about interactive systems and computers is, however, that they work error-free. Communicating uncertainty is therefore not trivial. Not communicating uncertainty could lead to users losing trust in a system because they assume it does not work correctly. Communicating uncertainty may, however, lead to users not even starting to use a system as they believe that it does not work properly.

Multiple studies in meteorology and psychology have shown that humans appreciate getting information about uncertainty and make better decisions when knowing about the uncertainty of presented data. These findings are, however, yet to be transferred to interactive systems. Additionally, an interactive system includes much more steps than looking at the data, for example the system has to handle user input. To show uncertainty in the output, the uncertainty in the user input has to be quantified. Current interactive systems mostly lack the opportunity for users to specify that they are uncertain about their input. They provide standard text or number fields for entering data. One example is an application for tracking calorie intake, where users have to enter the weight of their food in grams. Most users do not carry a scale and therefore guess the input values for the standard number fields. This falsifies the output as the system treats the values as reliable.

In addition to unavailable input options for users dealing with uncertain input, there are also no guidelines on when to use which visualizations and how this influences the usage behavior in the context of an interactive system. Additionally, it has to be explored how users can be supported with the correct interpretation of uncertain data. These are some of the core challenges that have to be tackled to support designers and developers in quantifying and displaying uncertain data in the best possible way for their target audience. Interactive systems dealing with these challenges could support users in making better decisions under uncertainty. The outlined core challenges therefore motivated our research questions presented in the next section.

1.1 Research Questions

Humans do not only deal with uncertainty in everyday life, but also more and more with uncertainty in interactive systems. Therefore, it is important to understand how users deal with uncertainty and how presented uncertainty in interactive systems affects users. New insights from research can support developers and designers in effectively quantifying and displaying uncertainty.

This thesis presents an exploration of five high-level research questions (RQs) to further increase the understanding of how developers, designers, and HCI researchers can make uncertain data understandable to the general public (see

Table 1.1: Summary of the five high-level research questions addressed in this thesis. More detailed questions are outlined in the respective chapters.

	Research Question (RQ)	**Chapter**
RQ1	What can we learn from the current usage of simulations?	Chapter 4
RQ2	What are the sources of uncertainty in interactive systems?	Chapter 5
RQ3	What input controls are suitable for uncertain input?	Chapter 6
RQ4	What visualizations are suitable for uncertain output?	Chapter 7
RQ5	How do people interpret uncertain data?	Chapter 8

Table 1.1). To answer these questions we built and evaluated prototypes with the help of empirical methods from HCI. We embed our research results in the application scenario of end-user simulations, which we define as simulations used by the general public. A simulation tool including a set of input and output methods is presented at the end of this thesis to foster future research.

Before exploring the topic of uncertainty in interactive systems and building concrete prototypes, we decided to ask experts and laymen how they currently use simulations to understand what we could learn from their usage behavior (RQ1). The findings served as foundation for this thesis. Additionally, we needed to identify the sources of all possible types of uncertainty in interactive systems to paint the full picture of how to deal with uncertainty (RQ2). In the following part of the thesis, we provide explorations for the suitability of different input methods (RQ3), output methods (RQ4), and users' interpretations of uncertain data (RQ5).

The next section of this introduction describes the methodology used in this thesis to address the outlined research questions.

1.2 Methodology

How to quantify and display uncertainty in interactive systems is a new and mostly unexplored field of research. However, no new technologies are needed to do research in this field. Due to the lack of design guidelines for how to quantify and display uncertainty in interactive systems, we used a bottom-up approach. We first started by observing current usage and identifying sources of uncertainty in interactive systems. Based on our findings, we explored the influence and suitability of different options that can be achieved with the current technology. Over three years we designed non-functional prototypes and implemented functional prototypes in projects of different scales and evaluated them with potential users.

1.2.1 Literature Analysis

We reviewed related literature about uncertainty and end-user simulation tools. We divided this literature into two parts extracting background information such as definitions, and classifications. Based on our literature analysis, we additionally enhanced the General Interaction Framework [Dix, 2009] with potential sources of uncertainty in interactive systems.

1.2.2 Requirement Analysis

As a foundation for this thesis, we aimed to learn from the current usage of simulations by both experts and non-experts. We conducted paper and online questionnaires, as well as a diary study, focus groups, and design workshops. We identified common usage patterns of simulation tools and key requirements for building an end-user simulation tool.

1.2.3 Prototypes

To understand the suitability and the possibilities of quantifying, displaying, and interpreting uncertain data in the context of interactive systems, we designed a range of non-functional prototypes, such as digital sketches and 3-D printed prototypes. Most of these prototypes were developed in a user-centered design process. We used interviews, focus groups, and online surveys to iterate on our design and identify potentially promising candidates. Based on the results, we implemented web-based study environments, a Facebook game, and Android applications.

1.2.4 Evaluation

We evaluated all our prototypes in lab studies or in-the-wild studies. Users either had to use the prototype in a lab environment or could, for example, play a

Facebook game or install an Android application to use it. We then asked them for their opinions by using standardized questionnaires, questionnaires with Likert scales, interviews, and qualitative feedback.

1.3 Research Context

The research that lead to this thesis was conducted at the University of Stuttgart in the HCI group over a course of three years. It was additionally part of a project funded in the Cluster of Excellence in Simulation Technology at the University of Stuttgart. Collaborations with internal and external colleagues inspired us to do this research.

Cluster of Excellence in Simulation Technology
The Cluster of Excellence in Simulation Technology at the University of Stuttgart offers a unique environment of interdisciplinary project networks. This thesis was conducted in the ongoing exchange of knowledge in the project network called “Reflexion and Contextualization”. The mid-term presentation of this thesis was accompanied by Prof. Dr. rer. pol. Dipl.-Ing. Meike Tilebein from the Institute for Diversity Studies in Engineering. Additionally, a publication in the context of this thesis was presented on a German conference on Economic and Social Cybernetics in 2014 [Greis, 2014]. The researchers of the clusters also participated in our focus groups, questionnaires, and studies as simulation experts and gave us feedback on the developed simulation tool.

Human-Computer Interaction Group, University of Stuttgart
The HCI group at the University of Stuttgart includes researchers with a broad range of knowledge. Collaborations with Passant El.Agroudy, Jakob Karolus, Hyunyoung Kim, Alexandra Voit, Dr. Tonja Machulla, Dr. Paweł W. Woźniak, Dr. Céline Coutrix, Jun.-Prof. Dr. Niels Henze and Prof. Dr. Albrecht Schmidt led to multiple publications [Greis et al., 2015, 2016, 2017a,d] and submissions currently under review, which are all in the scope of this thesis. Of particular success were two papers. The first, written in cooperation with Tonja Machulla won an Honorable Mention Award at the CHI’17 conference [Greis et al., 2017a]. The second, written in cooperation with students and Niels Henze won the ACM Best Student Paper Award at the EICS’17 conference [Greis et al., 2017d].

External Collaborations
In the context of the Cluster of Excellence in Simulation Technology, I was enabled to join the Living Mobile Group at the MIT Media Lab under the supervision of Chris Schmandt for three months. This collaboration led to a late

breaking work publication at MobileHCI'17 [Greis et al., 2017b]. The networking on conferences also lead to a workshop called “Designing for Uncertainty in HCI", which was accepted and conducted at the CHI'17 conference [Greis et al., 2017c]. Co-organizers of this workshop were Jessica Hullman (University of Washington), Matthew Kay (University of Michigan), Michael Correll (University of Washington), and Orit Shaer (Wellesley College).

1.4 Contributions

This thesis has three main contributions to the research about uncertainty in HCI:

1. We identify and present user requirements and sources of uncertainty in interactive systems.
2. We explore possible input methods, output methods, and interpretation strategies in the context of interactive systems.
3. We provide a simulation tool for future exploration of methods to enter and display uncertainty.

In the following, we present more details for each of the main contributions.

1.4.1 Requirements and Sources of Uncertainty

Based on a literature review and observations of users, we identified important requirements and sources of uncertainty. The related implications are the foundation of this thesis and the foundation of future work in the field. The enhanced General Interaction Framework includes possible sources of uncertainty that have to be considered and quantified when developing interactive systems. We identified three main areas of interest for the HCI research community: the input, the output, and the interpretation of uncertainty in the context of interactive systems.

1.4.2 Explorations of the Design Space

With the help of a series of non-functional and function prototypes, we explored the key aspects identified based on the sources of uncertainty: the input, the output, and the interpretation of uncertainty in the context of interactive systems. We contribute novel methods for the explicit and implicit input of uncertainty. We explored how to design standard input controls, specialized slider controls, and tangible input controls for entering uncertain data. For each exploration, we contribute a most promising design. We additionally showed the feasibility of using behavioral and physiological measurements to implicitly capture uncertainty. On the output side, we identified a lack of communication of uncertainty in current mobile applications although our participants voiced a preference for it. We therefore contribute concrete designs that improve the communication of uncertainty for activity tracking data. We further contribute a classification of representations based on the amount of uncertainty information included in a representation and findings on how different amounts of uncertain information impact decision-making. Regarding the interpretation, our research indicates that user-made predictions can improve users' reasoning about uncertainty and predictions. We also contribute concrete design recommendations for showing conflicting information to either increase users' confidence or improve their internal models of reasoning about conflicting data.

1.4.3 Simulation Tool

We further contribute a simulation tool that includes the novel input and output methods developed and compared in our studies. The tool serves as a platform for future research in the area of uncertainty in HCI to gather insights into how users might use new input controls and visualizations.

1.5 Thesis Overview

The body of this thesis consists of four parts, which contain eleven chapters. After this part *Introduction and Motivation*, the part *Foundation* presents an in-depth literature review of related work about uncertainty in different research areas. Literature only relevant for one chapter is discussed in the introduction of the respective chapter. Additionally, it contains research on current simulation usage and key requirements for an end-user simulation tool. This is followed by the main part of this thesis; *Uncertainty in Interactive Systems*. This part contains the identification of sources of uncertainty in interactive systems and explorations of input methods, output methods, and the interpretation of uncertainty. The thesis closes with the *Conclusion and Future Work*, which introduces the simulation tool, summarizes and discusses the research contributions of the whole thesis, and presents future work.

Part II: Foundations

This part creates the foundations for the thesis. It contains background information, an in-depth literature review and smaller research probes for identifying key requirements for an end-user simulation tool.

Chapter 2: Background

In this chapter, we introduce and define the terms *uncertainty* and *simulation*. We additionally provide an overview of classifications and taxonomies for uncertainty from different research areas including information visualization, medical visualization, and Geographic Information Science (GIScience). The chapter provides the background knowledge for the topics discussed in this thesis.

Chapter 3: Related Work

Chapter 3 introduces related work relevant for all chapters of this thesis. In the first section, we highlight the importance and the challenges of uncertainty research including perspectives from different research areas. The following section introduces work on textual and iconic communication of uncertainty and discusses advantages and disadvantages of linguistic and numerical communication of uncertainty. We further introduce related work that compares numerical, verbal, and iconic representations for uncertainty. The third section briefly introduces uncertainty visualization in the visualization community, then focuses on the visualization of uncertainty for the general public. We mainly focus on four subtopics which are visualizations in general, visualizations for weather forecasting, visualizations for geographic data, and the specific use of uncertainty visualization in HCI. We introduce different visualization possibilities and discuss their advantages and disadvantages. In the fourth section, we discuss the interpretation of uncertainty data focusing on the problems of judgements under uncertainty, decision-making and concepts such as confidence and trust. In the next section, we discuss the topic of end-user simulations and present some related work on simulation tools, and in the last section we summarize the implications of the presented related work for this thesis.

Chapter 4: Understanding Simulation Users

In Chapter 4, we present smaller research probes on the current usage of simulations by simulation experts and non-experts. In the first section, we provide definitions for these terms to distinguish between user groups, then present the method and results of an online survey and a paper questionnaire conducted with simulation experts. Furthermore, we present the methods and results of a diary study, focus groups, and design workshops conducted with non-expert, in which we explore the current usage of simulations in everyday life, potential use cases, and the potential future usage of simulations and summarize our insights for the development of an end-user simulation tool.

Part III: Uncertainty in Interactive Systems

Part III contains the main content of this thesis. It contains four chapters that focus on one high-level research question each. The identification of sources of uncertainty in Chapter 5 is based on related work. All other chapters contain different explorations of uncertainty in interactive systems. All explorations contain an evaluation.

Chapter 5: Sources of Uncertainty

In this chapter, we first introduce the General Interaction Framework by Dix [2009], then enhance the General Interaction Framework to contain sources of uncertainty in interactive systems. We identify these sources based on related work, and provide an overview on the implications for HCI research.

Chapter 6: Input Methods

Chapter 6 contains four explorations of input methods for entering uncertainty. In the first section, we explore how to enhance standard input controls to allow users to enter additional uncertainty. Based on a taxonomy for common input controls and methods for entering uncertainty, we built a set of sketches that were refined in a pre-study. We then conducted an evaluation in the lab to identify the most promising user interface. In the second section, we present our designs for probability distribution sliders. These are specialized slider controls that offer more transparency and flexibility than standard input controls. We evaluated the sliders in an online survey and a lab study and provide implications for their use. The third section introduces designs for tangible shape-changing input controls.

We conducted focus groups to identify the most promising design Split Slider and evaluated this design in a lab study. We further discuss why shape-changing input controls are a promising area for the input of uncertainty. The next section introduces physiological sensing as an implicit method for measuring uncertainty. We present our methods and evaluation in the lab. The last section of the chapter summarizes our insights into quantifying uncertainty in user input.

Chapter 7: Output Methods

To understand the state of output methods for uncertainty communication, we analyzed current mobile applications on how they communicate uncertain data. The first section of this chapter describes our analysis and contains the results of an online survey on users' expectations for uncertainty communication. We further designed three different graphical overview communicating uncertainty for activity tracking data that we introduce in the next section. We conducted an online survey to inspire our designs and then evaluated the designs in an in-the-wild study. In the third section, we introduce a classification of representations according to the amount of uncertainty information. We additionally present the results of an online survey and an in-the-wild evaluation of different visualizations for uncertainty to understand their influence on decision making. We complete the chapter with a section describing our insights on communicating uncertainty in interactive systems.

Chapter 8: Interpretation

In this chapter, we present three research probes on the interpretation of uncertain data. The first section presents an android application and an evaluation of the application which leveraged user-made predictions to teach users behavior patterns. In the second section, we present the method and results of an online survey and an in-the-wild study which compared different aggregation mechanisms for data from multiple weather source providers. In the following section, we present the methods and results for an experiment in the lab where participants had to indicate the true value for two conflicting sensor measurements. These measurements were depicted with different visualizations to learn about humans' internal models of aggregating conflicting data when using uncertainty visualization. We outline our insights of the chapter in the last section.

Part IV: Conclusion and Future Work

This part consists of three chapters that summarize and conclude the work presented in the previous chapters. Based on our research, we implemented a simulation tool for end-users. We additionally present the conclusion and potential directions for future research.

Chapter 9: Simulation Tool for End-Users

Based on the research presented in Part III, we present SimulaTE - a simulation tool for end-users. The tool includes input methods and output methods for uncertain data and was built based on the key requirements identified in Chapter 4. The tool should serve as a basis for future research on uncertainty in interactive systems.

Chapter 10: Conclusion

In this chapter, we summarize and present the main contributions and conclusions from this thesis. The discussions focuses on the research questions identified in the *Introduction* chapter of this thesis.

Chapter 11: Future Work

Chapter 11 focuses on future work. We identify and discuss potential directions for future research and follow-up projects.

II

Foundations

Chapter 2

Background

In this chapter, we present background information, definitions, and classifications relevant for this thesis. We focus on definitions and classifications of uncertainty, and simulation as a specific application area where uncertainty plays an important role.

2.1 Uncertainty

Uncertainty quantification and visualization is an object of research in many different domains. As most research about uncertainty is domain-specific, manifold definitions and different uses of the term *uncertainty* exist. Additionally, the distinction between uncertainty and related terms such as *data quality*, *reliability*, *accuracy*, and *error* mostly remains unclear [MacEachren et al., 2005].

Pang et al. [1997] define *uncertainty* “to include statistical variation or spread, error and differences, minimum-maximum range values, noisy, or missing data.”, covering all possible types and sources of uncertainty. This definition is widely used in the visualization community. Other definitions are similarly open such as that by Bonneau et al. [2014], who refer to uncertainty as “the lack of information”. Similar to Potter et al. [2012], they differentiate between epistemic and aleatoric uncertainty, in which epistemic uncertainty could, for example, be introduced by wrong measurements or models, while aleatoric uncertainty is inherent random uncertainty that can be calculated statistically. The National Institute of Standards and Technology (NIST) guidelines also point to two types of measurement uncertainty [Taylor and Kuyatt, 1994]. Besides these, other very strict definitions exist. Gershon [1998] for example sees uncertainty as one singular factor of imperfect knowledge where the information is actually known but the user is unsure about its accuracy.

One example of how researchers disagree on the definition of uncertainty also manifests in the area of model-based decision support. While Walker et al. [2003] define *uncertainty* "as being any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system", Norton et al. [2006] highly disagree with this definition as they consider it to be problematic from a social science perspective.

In this thesis, we follow the definition of Pang et al. [1997] which includes all possible types and sources of uncertainty. The main reason for following Pang's definition is that we are not interested in one specific type of uncertainty, but rather in a holistic view on uncertainty in interactive systems. We therefore want to incorporate as many understandings of uncertainty as possible.

2.2 Types and Sources of Uncertainty

In addition to multiple definitions of the term *uncertainty*, several taxonomies and typologies in different domains and with different foci exist. They either refer to the sources of uncertainty, classification of visualization methods, dimensions, or types of uncertainty.

Skeels et al. [2008] identified types of uncertainty commonly discussed in literature. They found three levels of uncertainty: measurement precision, completeness, and inference. They additionally found two categories that span levels: disagreement and credibility. Disagreement refers to data either measured multiple times or provided by disagreeing sources. Credibility, however, refers to a data source producing unreliable data. In this thesis, we do not focus on any specific type of uncertainty as all of the types identified by Skeels et al. might be present in interactive systems and therefore relevant for the HCI community.

In their work on advanced navigation systems, Andre and Cutler [1998] identified three dimensions of uncertainty: accuracy (e.g. biases in measurements), precision (e.g. data measured with a low level of precision), and time (e.g. delivering information with a time lag). They developed these dimensions in the application scenario of information displays for pilots. Their dimensions mostly correlate to some of the types of uncertainty identified by Skeels et al. [2008]; however, they are transferred to a specific application scenario and contain the time lag as a very specific dimension of uncertainty in information displays.

Focusing on uncertainty visualization, Pang et al. [1997] reviewed sources of uncertainty in the context of the visualization pipeline. They further classified uncertainty visualization methods based on their value, location, data, extent, visualization extent, and axis mapping. Potter et al. [2012] refined existing classifications and focused on only two qualities: data dimension and data uncertainty dimension, providing examples for all combinations. Although these classifications provide a good overview of uncertainty visualization they are rather complex and apply to visualizations used by visualization experts. They might not be as suitable for the HCI community.

Ristovski et al. [2014] present a taxonomy of uncertainty types in medical visualization based on spatial location, dimensionality, type of events, and sources. They identified sources of uncertainty such as noise, discretization, interpolation, or human interpretation. Of these, human interpretation as a source of uncertainty might be of particular relevance for HCI research.

Walker et al. [2003] describe their conceptual framework for understanding uncertainty in decision-making, proposing to distinguish between three dimension of uncertainty: location (where the uncertainty manifests in the model), level (from known to completely unknown), and the nature of uncertainty (type of uncertainty). The locations of uncertainty could also be described as sources of uncertainty. However, their work only focuses on the model, not on the role of the user in the decision-making process.

Further specialized typologies exist in other domains. One example is the typology developed by Thomson et al. [2005] for geospatial data and intelligence analysis. Some categories of their typology are related to the types of uncertainty discussed by Skeels et al. [2008]; however, their typology is much more detailed and contains a timing category which may refer to the time lag described by Andre and Cutler [1998]. One drawback of the typology is that it does not relate the identified categories to sources of uncertainty.

Although multiple taxonomies and typologies for uncertainty exist, there is no agreement on which of these taxonomies or typologies should be used in HCI. As none of them refers to a system with user input, we argue that a theoretical examination of sources of uncertainty in HCI is needed, which we provide in Chapter 5.

2.3 Simulation

Simulation is a very powerful technique used in different fields to explore the behavior of complex systems such as, for example, the flow of ground water, human walking behavior, and the world's climate. Due to its applicability to many problems, simulation is one of the most used techniques in research and management sciences [Law, 2015]. In this thesis, we use a definition of *simulation* following Shannon [1998]: Simulation is the development of a mathematical model imitating real systems to help to understand the system or develop management strategies for the system, because the adaption of the real system is, e.g., impossible or too expensive.

What has mainly hindered simulations so far from gaining more importance in everyday life, is that their execution takes a lot of computing time [Law, 2015]. However, computing power is steadily increasing; thus, the computation of simulations will increasingly become feasible on mobile phones and similar small personal devices.

Chapter 3

Related Work

Related work in many different research areas, although mainly in HCI and visualization is of interest for this thesis. In this chapter, we present related work about the importance of the research topic uncertainty itself, uncertainty visualization, uncertainty communication in HCI and beyond, and related work for one application scenario of uncertainty communication: end-user simulations.

Uncertainty visualization and communication to experts and the general public is relevant for many application scenarios such as ensemble weather forecasts [Gneiting and Raftery, 2005], context-aware systems [Antifakos et al., 2004; Lim and Dey, 2009], navigation systems and maps [Andre and Cutler, 1998; MacEachren, 1992], object classification [Bisantz et al., 2011], machine learning [Kay et al., 2015], data analysis [Ferreira et al., 2014], and information fusion [Riveiro, 2007]. Application scenarios can already be very concrete such as interactive tools to explore personal genomics [Shaer et al., 2016], public transport predictions [Kay et al., 2016], or the potential range of an electric car [Jung et al., 2015]. We argue that this makes uncertainty an interesting and important research topic as no general understanding or guidelines about uncertainty exist so far.

3.1 Importance of and Challenges for Uncertainty Research

As Couclelis [2003] stresses in her work, uncertainty is included in complex knowledge no matter what scientists do about it. She argues that scientists often present information as fact, when in reality it is affected by uncertainty. Uncertainty communication for the public is therefore very important as nowadays many people can provide information on the internet, where it is unclear on how certain or uncertain the sources are. Although Couclelis mainly refers to geographic information systems (GIS), we argue that this nowadays holds true for other systems as well. With the huge amount of information freely available on the internet, it is more difficult for the public to understand whether data sources are reliable, or whether presented information is uncertain.

We argue that uncertainty communication practice follows the same phases of risk communication described by Leiss [1996]. According to Leiss, in phase 1 uncertainty has to become manageable and quantifiable. In phase 2, the communication needs to gain attention and in phase 3, the responsibility of both dimensions has to become a part of the normal practice. Five years later, Lipkus et al. [2001] still identified the communication of uncertainty as one important issue in graphical risk communication that needs to be addressed. Even today, the communication and visualization of uncertainty is still a difficult problem with no clear solution at hand.

Brodlie et al. [2012] presented several reasons why it is so hard to deal with uncertainty in visualizations. First of all, uncertainty itself is very complex and can be presented with various different visualizations, for example as bounded data or probability distribution function. Uncertainty can be introduced at the modeling step and then adds up (e.g. through linearity in the computer hardware) and propagates through the visualization pipeline. In the visualization itself, uncertainty is often added as an additional dimension as it has to be visualized in addition to the deterministic values and often dominates certainty when for example the error bars are very long. In general, uncertainty also makes topics interdisciplinary as a lot of different groups of people have to work together to correctly quantify, propagate, and visualize uncertainty. This clarifies that the nature of the topic uncertainty itself turns it into a complex research area with many pitfalls. As HCI is an interdisciplinary field, uncertainty research should be included to understand how interactive systems are affected by uncertainty.

MacEachren et al. [2005] identified future research challenges for the visualization of uncertainty in geospatial data. Some of these are highly interesting for HCI, such as the identification of the components of uncertainty and how they relate to users' needs as well as understanding the usability of visualizations and tools that help capture, represent, or interact with uncertain data. The American Meteorological Society (AMS) [2008] outlined very similar challenges with probability information in weather forecasts. The challenges here also focused on the tools for communicating uncertainty information, mainly on how users can easily understand uncertainty information and how they can be supported in actually interpreting such information. As these challenges span different domains and reach far into HCI, we therefore see uncertainty communication in different domains as promising application areas for the field.

Boukhelifa and Duke [2009] summarized the challenges in three points that are crucial for the visualization of uncertain data: First, the quality and the scope of the uncertain data has to be good, which automatically includes the knowledge of sources of uncertainty and the correct quantification of uncertainty contained in the data. Second, the limited confidence in the data is a major problem. And third, the visualization itself can confuse people to make wrong assumptions or interpretations of the data. HCI is a discipline that can help tackle and probably solve some of the recent problems around uncertainty communication both to experts and the general public.

3.2 Textual and Iconic Communication of Uncertainty

Uncertainty communication can have various forms and consequences. Systems communicating uncertainty might fail to convince users that they do actually work. However, in automated systems not communicating uncertainty, the data might be perceived as more reliable or accurate by users than it really is. When this has impact on decisions, users may lose trust in an automated system not communicating uncertainty [Andre and Cutler, 1998]. This indicates that interactive systems need to find the optimal compromise in presenting uncertainty information without overwhelming or confusing users.

For the textual communication of uncertain data, either numerical or linguistic expressions or a combination of both can be used. In the following, we present research using both methods and additionally icons to communicate uncertainty.

3.2.1 Linguistic Communication

Early researchers into this question assigned probabilities to linguistic expressions such as “almost certain” or “probable” [Kent, 1964], but humans have different perceptions of terms such as “low risk” and “low uncertainty” [Wallsten et al., 1986]. Budescu et al. [2009] conducted a study where participants assigned numerical values to linguistic expressions. The judgements varied considerably across subjects, and the authors recommended specific guidelines to make linguistic communication of uncertainty easier: First, there needs to be a differentiation between the uncertainty of the specific event and general ambiguity in the description (e.g., a big storm). Second, specifying the sources of the uncertainty stating their nature and magnitude could help people to better understand the information. Third, linguistic and numerical communication together seems to be most promising and fourth, ranges of uncertainty classes should not be strict but adapt depending on the event. These guidelines could also be applied to interactive systems using linguistic expressions to communicate uncertainty.

Another factor that plays a role for the linguistic communication of uncertainty is framing. Spiegelhalter et al. [2011] present examples of the influence of positive and negative framing on the understanding of uncertainty information.

3.2.2 Numerical Communication

As linguistic communication, communicating risk in numerical ways also has disadvantages. As Lipkus et al. [2001] found in their conducted studies, even one-fifth of their highly educated participants were not able to answer easy numerical questions such as “Which represents the lagrer risk: 1 %, 5 %, or 10 %?” This may also have impact on uncertainty communication as uncertainty is often communicated as a probability or percentage. It is important to better understand how people interpret such values. The authors propose training as one option for improving humans’ understanding of numerical expressions and probabilities.

One specific problem with weather forecasts is the communication of the probability of precipitation which is often interpreted wrongly [Gigerenzer et al., 2005]. The main problem is that the reference class for the probability is missing, thus forecast users do not know what the probability refers to. For a probability of precipitation of 30 %, they will for example often assume that it will rain 30 % of the time or in 30 % of the area instead of understanding that it will rain in 3 out of 10 days that are like the forecasted day.

Gigerenzer and Hoffrage [1995] conducted two studies comparing frequency and probability formats for communicating uncertainty in well-known problems such as the mammography problem. They found that the frequency formats improved participants' Bayesian reasoning. They also proposed to better teach students with different statistics education about how to convert probability into frequency formats to increase the understanding. However, this finding seems to be dependent on the task and problem presented to participants. Frequency formats seem to be less optimal to communicate uncertainty in weather forecasts and rather confused participants [Joslyn and Nichols, 2009]. Weather forecast users actually seem to prefer numerical probabilities in weather forecasts and are able to understand them [Murphy et al., 1980]. Whether to use frequency formats or numerical probabilities should therefore carefully be evaluated based on the application area and the presented information.

One important aspect of uncertainty communication includes the formulation of uncertainty. Joslyn et al. [2009] found that mismatches between a given piece of information and a given task could lead to confusion, e.g., issuing a wind warning if the wind exceeds a certain limit, but the information provided is the probability that the wind is less than the specific limit. They conclude that the provided uncertainty information has to be in line with the goal and the task.

Teigen and Jørgensen [2005] combine two ways of expressing uncertainty: probabilistic modifier (such as 90 % probable) and interval estimates (e.g., 3 to 4 weeks) to have intervals with probability estimates. They ran multiple experiments to study how credible such confidence intervals were. They showed that participants estimated the confidence very differently than the actual confidence of the intervals and that the estimated confidence was dependent on the interval size.

3.2.3 Iconic Communication

Bisantz et al. [2011] additionally explored the use of transparent icons in situations of decision-making related to object classification. In one study they compared three different visualizations: a solid icon of the most probable classification, a transparent icon of the most probable classification where transparency matched probability, and a transparent icon of a missile that was the object for participants to react on. They found that a toggle mode allowing participants to toggle between the different representations worked best. For interactive systems, it might actually be possible to support multiple representations and provide a toggle functionality for users.

3.2.4 Comparisons

Bisantz et al. [2005] compared numerical, verbal, and iconic representations for uncertainty information. All of these worked well to communicate uncertainty information. More fine-grained steps in the representations and icons helped participants to make more conformed decisions and take fewer risks. However, this might also depend on the given task and information.

Lipkus [2007] identified research directions for the verbal, numerical, and graphical communication of risk. For graphical displays, more work needs to focus on understanding the impact of these visualizations on risk perception and the actual interpretation of the information by users. One interesting aspect is that the author proposes to engage the users interactively to explore the presented data which might sharpen their mental representation of the data and the associated risk. This indicates that interactive systems might be a good tool to educate users about uncertainty.

3.3 Uncertainty Visualization

Uncertainty visualization is a well explored topic in the visualization community. Research focuses on different visualizations, user groups, and data dimensions. Additionally, many other research fields evaluate uncertainty visualizations for specific application scenarios. In this thesis, we mainly focus on work relevant for communicating uncertainty to the general public. However techniques such as shading [Jackson, 2008], animation [Ehlschlaeger et al., 1997], simulations with pixel mixing [Hengl and Toomanian, 2006], summary plots [Potter et al., 2010], or glyphs [Wittenbrink et al., 1996] and other complex methods for uncertainty visualization have been explored and used in the visualization community.

In the following, we present work from different research fields such as visualization, GIScience, weather forecasting, and HCI.

3.3.1 Visualization in General

Olston and Mackinlay [2002] propose to use different representations for two different types of uncertainty: statistical and bounded uncertainty. For statistical uncertainty, they propose to use error bars, whilst for bounded uncertainty, they propose to use ambiguation. However, as they did not evaluate the visualizations with users, it is difficult to assess whether non-experts in statistics would be able to understand the differences between them. Additionally, ambiguation might often be used for confidence intervals and then lead to confusion about whether statistical or bounded uncertainty is being visualized.

Tak and Toet [2014] conducted a study with non-experts comparing seven variations of a line chart. The uncertainty information was included by adding additional components such as solid borders, a gradient or a confidence interval to the line chart. The results showed that the representation influenced how certainty was perceived by the participants. However, all representations communicated uncertainty as participants interpreted the borders with less certainty than the middle of the representation. Most suitable according to their findings were dashed borders, random lines, and gradients as participants' answers on the certainty fitted a normal distribution quite well. This work provides interesting insights into how suitable these techniques are to match a given certainty; however, it is unclear how this affects users' decision-making and whether they can use the representations for reasoning.

Error bars are a common tool to communicate uncertainty for different chart types, mainly used in the scientific or expert context. Sanyal et al. [2009] ran a user study to compare four uncertainty visualizations for 1D and 2D datasets: error bars, scaled glyphs, colored glyphs, and color-mapping on the data surface. Participants had to for example search or count the most uncertain or certain data points. The study did not lead to clear results, but error bars performed worst. The authors speculate that the effectiveness of the visualizations highly depends on the particular task. Correll and Gleicher [2014] also found that bar charts with error bars may lead to reasoning biases such as the "within-the-bar bias". A value in the bar is seen as more likely than a value outside the bar although they may have the same probability. Visually symmetric and continuous plots such as gradient or violin plots seem more promising in avoiding such biases. Error bars are additionally misleading as they can either stand for the standard error, the standard deviation, or a confidence interval [Belia et al., 2005]. Belia et al. also found that researchers do have problems in relating error bars to statistical significance and do not understand the impact of a with-in- or in-between-group design. We can conclude that although error bars are widely used, they are not the optimal tool for communicating uncertainty even in a scientific context. Uncertainty visualization techniques that are already adopted by a community do not necessarily lead to a better understanding of presented data. New ways of communicating uncertainty are still needed, because they may outperform established methods.

Gschwandtnei et al. [2016] compared six different representations for communicating the uncertainty of temporal data: three representations showing statistical uncertainty and three showing bounded uncertainty. However, in contrast to Olston and Mackinlay [2002], they define error bars as bounded uncertainty instead of statistical uncertainty. Participants in the study had to specify start and end times, interval durations, probabilities, and personal preferences in relation to the representations. The results show that ambiguation or error bars work best for start and end times as well as minimum and maximum times; however, for the probability aspect a gradient worked best although participants did not prefer the gradient representation. This shows that representations encoding bounded uncertainty are suitable for tasks that do not require to use probability values. If, however, the probability of the points in time plays an important role, then more complex representations have to be used.

3.3.2 Geographic Information Science

MacEachren [1992] proposed different examples to show uncertainty on maps, for example, by reducing the clarity, adding fog, or reducing the resolution of a map part. The information could either be shown on a second separate map, sequentially presented (e.g. by toggling the two maps), or by incorporating all the information in a bivariate map. The author stresses the importance of evaluating such alternatives to identify the most promising candidates. Empirical research methods from HCI could be used to evaluate interactive interfaces showing maps with uncertain data.

Aerts et al. [2003] conducted a web-based survey with maps that showed uncertainty by varying the color intensity of the map. They provided two different versions: one version in which participants saw the model result next to a map showing the uncertainty and a second version in which participants were able to toggle between the model result and the uncertainty information. They found that all participants - experts and novices - did not find the display of uncertainty more complex, but instead found that embedding uncertainty rather clarifies. This is, however, only a preference of people to like uncertainty displays, but does not yet allow any conclusions on how uncertainty information on maps influences decision making.

MacEachren et al. [2012] also explored how to display uncertainty for single objects (e.g. locations on a map). They evaluated a broad range of visual variables such as fuzziness, location, and saturation. Participants preferred fuzziness and location, while saturation was ranked very low. They also tested iconic signs, which can sometimes be more intuitive than visual variables if the metaphor of the icon is easy to grasp. The work provides a good overview of visual variables and could be used to explore these in different application areas.

3.3.3 Weather Forecasting

The National Research Council (NRC) already raised the need for uncertainty communication in weather forecasting in 2006: “Uncertainty is thus a fundamental characteristic of weather, seasonal climate, and hydrological prediction, and no forecast is complete without a description of its uncertainty.” [National Research Council, 2006]. Therefore much work on communicating uncertainty to the public has been conducted in the field of weather forecasts.

Joslyn and Savelli [2010] ran a survey that showed that weather forecast users are aware of the uncertainty and factors (such as the time) that increase the uncertainty in a forecast. However, they judge the uncertainty very differently as they seem to have high variations and unjustified biases. The authors suspect that humans may have a bias towards a climatological norm, which results in them perceiving extremely high or low values as less probable than what they consider to be normal values. Explicit uncertainty visualization could help to better convey the information and reduce biases.

Ibrekk and Morgan [1987] looked at a greater variety of representations for communicating uncertainty to non-experts in the weather context. They compared representations such as confidence intervals, box plots, pie charts, gradients, and probability density functions. In the presented study, participants had to solve concrete tasks such as estimating the mean of snowfall, the probability that a specific amount of snowfall had been exceeded, or the probability that the snowfall lies within a certain range. Based on their results, the authors propose to use a probability density function in combination with a cumulative probability density function.

Frick and Hegg [2011] conducted a longitudinal study with a visualization platform for meteorological and hydrological information. The platform did not provide an interpreted weather forecast, but instead the outputs of multiple ensemble models. They found that users of the platform preferred to have detailed information about uncertainty to make their own judgements, but that new users might need some time to adapt and get comfortable with the system. Nevertheless, the display of the ensemble models increased users’ confidence and may also support decision-making. One challenge that the authors however identified, is that users might see uncertainty as a shortcoming rather than a feature.

In contrast to understanding the interpretation of weather forecasts by the general public, other studies worked with meteorologists to understand their needs and opinions on weather forecasts. Morrow et al. [2008] ran focus groups with broadcast meteorologists to understand their view. They raised concerns on how uncertainty communication can fit into the limited time span of broadcasts and feared that this might impact on the competition with other broadcasters. The broadcast meteorologists rather were convinced that the general public wanted definitive answers and that uncertainty could most probably be communicated through the delivery of the forecast without using numerical expressions. In contrast to their view, forecasters are convinced that forecasts should be displayed as probabilities [Murphy and Winkler, 1971]. Here, it is important to note that different groups and stakeholders might disagree on whether uncertainty should be presented in a specific application scenario.

Pappenberger et al. [2013] ran a study with expert users of Hydrological Ensemble Prediction Systems (HEPS). Their participants stated that uncertainty is often not released to the general public as it is unclear how to best communicate the uncertainty to make it understandable for non-experts. However, they agreed that for experts, it is very important to have the information about the uncertainty.

3.3.4 Human-Computer Interaction

Kay et al. [2016] explored how to present public transport predictions on mobile devices. They compared density plots, stripe plots and dotplots, which are a novel discrete representation that can be used instead of continuous density plots. In a user study, they found that dotplots improved probability estimates and confidence of participants. The dotplot seems to be a promising alternative to probability density plots for communicating uncertainty to non-experts.

A special application area for uncertainty communication is the exploration of personal genomics data. The area is special as the uncertainty does not come from the data itself, but from its interpretation and the implications drawn from new research in the area [Shaer et al., 2017]. Shaer et al. developed GenomiX, which allows users to interact with their personal genomics data. They use three categories to reflect the uncertainty from “uncertain” to “well-established”. The data is then shown in a matrix of the certainty and the severity of the health effect. Additional color-coding, size, and border of the displayed gene variants are used to communicate additional properties. This shows that HCI has to deal with different types and sources of uncertainty depending on the application scenario.

3.4 Interpretation of Uncertain Data

Many problems in communicating uncertain data arise due to missing knowledge or misinterpretation of the presented data. In the following, we present related work focusing on the understanding of uncertain data, the influence of communicating uncertain data on decision-making, and how presenting uncertainty information impacts confidence and trust.

3.4.1 Problems of Understanding

Very early on, Tversky and Kahneman [1975] identified a set of biases and heuristics that people have when making judgements under uncertainty. Humans are, for example, insensitive to the prior probability of outcomes and to the size of the sample. The authors also argue that humans struggle to understand regression and tend to believe what is imaginable by them. Such biases and heuristics have to be taken into account when communicating uncertain data.

Besides understanding statistics and uncertainty information, humans also have problems in understanding accumulation in stock-flow systems. Sweeney and Sterman [2000] conducted experiments to examine the general system thinking ability of participants focusing on stocks and flows (e.g., inflow, outflow, and stock of water in a bathtub). Well educated participants had huge problems in understanding the relationships between the net inflow and the slope of the stock. Humans probably match the behavior of the stock to the pattern of the flow which even seems to be independent of other factors such as knowledge and motivation [Cronin et al., 2009].

3.4.2 Decision-Making under Uncertainty

In Psychology, multiple studies have showed that providing uncertainty information improves decision making. Roulston et al. [2006] conducted a study where participants had to play a game in which they needed to salt the road based on an overnight minimum temperature forecast. Groups of participants received different information to play the game: a point estimate, a point estimate and the standard error, or an additional probability of freezing. The results indicate that uncertainty information (in this case the standard error) helped participants to increase profit and make better decisions. Joslyn and LeClerc [2012] ran a very similar study where people performed significantly better on all measures when they had uncertainty information to decide whether to salt a road or not. Providing a decision aid without providing uncertainty information did not help participants to make better decisions. In contrast, forecasts providing uncertainty information seemed less wrong for participants and overall more reliable. This shows that showing uncertainty can actually increase trust and has positive consequences. Additionally, people like to make their own decisions and do not necessarily follow decision aids but rather follow their own judgement, especially if forecasts seem to be wrong.

Nadav-Greenberg and Joslyn [2009] ran two experiments to examine whether people make better decisions if confronted with uncertainty information in weather forecasts. They used different verbal, numerical, and one visual representation to present the forecast. The results showed that presenting uncertainty does not lead to perfect, but more optimal decisions than without uncertainty information. However, more information was not necessarily better. Depending on the format of presentation, people performed better or worse. In the experiment, participants also received feedback, so it might require some training to learn how to use the additional information.

Savelli and Joslyn [2013] conducted three experiments in which they showed that forecasts using predictive intervals lead to better decisions as they support the identification of unreliable forecasts and help participants to gain a better understanding for the range of the outcome. In their experiments, they compared a deterministic forecast with a verbal presentation, a plus/minus presentation, and a bracket representation. In one experiment, they also added a gradient visualization. One main finding was that participants interpreted some information as diurnal fluctuation instead of predictive intervals. This indicates that predictive intervals might be better communicated in text. However, the visualization might just have been too close to what users normally know to be used to communicate diurnal fluctuation and therefore was mistaken.

Ramos et al. [2013] conducted an experiment where participants acted as decision-makers in a flood forecast scenario. They had to make a decision on whether to open a flood gate to avoid flooding a town. The flood forecast was provided including different levels of uncertainty information. The authors found that uncertainty information increased participants' optimal decisions and lead to a smaller variance in all decisions. Additionally, the higher the uncertainty was, the more participants avoided taking a risk and opened the gate.

Roulston and Kaplan [2009] present another study where participants got a five-day weather forecast to decide which one of two criteria would most likely occur during the next days. One group had to make the decision based on a point estimate while the other received uncertainty information. The latter group performed better than the first group, but did not need significantly more time for their decision. The authors speculate that participants without uncertainty information may have made it up in their heads as they were probably aware about the uncertainty of the forecast.

All studies presented in this subsection show that participants made better decisions when uncertainty information was presented to them. This indicates that uncertainty should be communicated in interactive systems to support users in making most optimal decisions.

3.4.3 Confidence and Trust

Morss et al. [2008] conducted a nationwide survey of the U.S. public about forecast uncertainty. They found that forecast users were well aware of the uncertain nature of forecasts and expect forecast errors, and stated a preference for forecasts explicitly communicating uncertainty. Users' confidence in forecasts is mainly influenced by two factors: First, the time span of the forecast, as users are less confident in long-term forecasts; and second, the forecasted information, for example users generally seem more confident in temperature forecasts than in the probability of precipitation. In a further survey including scenarios in which participants had to make decisions on taking proactive steps or not, Morss et al. [2010] found that participants use different criteria to determine whether proactive steps are necessary or not. The authors conclude that forecasts should not show concrete decision criteria to allow users to use their own criteria to make individual judgements.

Very similar is the finding of Jung et al. [2015] that displaying uncertainty in an electric car will reduce users range anxiety. Although single numbers are easier to read and understand, disguised uncertainty can impact on the user experience and increase anxiety. The same principle applies for bathroom scales where gaps in the understanding of the users about the accuracy and fluctuations of body weight might decrease the trust in scales. Kay et al. [2013] therefore proposed to redesign the interface of scales instead of focusing on accuracy to consider the uncertainty in the measurement and due to fluctuations.

Regarding context-aware systems, it is not really clear whether uncertainty information improves user experience and helps to generate trust. Antifakos et al. [2004] found that displaying uncertainty can actually have positive effects on task performance and that uncertainty information helps users to better understand the system state and how well it works. Additionally, Lim and Dey [2009] identified certainty as important information that should be provided to the users of context-aware systems although it may depend on the context or kind of information provided. In a further piece of work, Lim and Dey [2011] conducted a survey with usage scenarios for context-aware systems. They found that showing uncertainty for applications with high certainty increases users trust and makes it easier for them to forgive errors. However, if the uncertainty of an application is too high, users might lose their trust in the application and perceive it as not good enough to be used. Rukzio et al. [2006] showed uncertainty in a user-study for an application which autofilled forms on websites. Uncertain fields were highlighted by colors. Participants did not trust the visualized uncertainty and checked each field for correctness anyway instead on taking the visualized probability into account. Thus, in this specific use case displaying uncertainty was not advantageous. For context-aware applications, the display of uncertainty might therefore heavily depend on the importance of the task and the goal of the users.

In machine learning, classifier accuracy is a common measure. Kay et al. [2015] developed an evaluation method for the acceptability of the accuracy. This is an important step towards creating machine learning algorithms that increase confidence and trust of users.

3.5 Simulation Tools for Non-Experts

For educational purpose, simulation tools for non-experts already exist. These tools were especially created to allow children or students to play with simulations. One of the first tools in the educational context was the Alternate Reality Kit [Smith, 1986] which allowed students to build physics simulations. Another early tool was Playground [Fenton and Beck, 1989] that allows children to add rules to graphical objects to construct simulations. Most of the existing simulation tools for non-experts use agent-based modeling approaches [Railsback and Grimm, 2012] and try to substitute or simplify programming. KidSim [Cypher and Smith, 1995; Smith et al., 1994] uses the technique of programming by example/demonstration, whilst users of StarLogo [Resnick, 1996] have to attach puzzle pieces to each other to create a simulation. NetLogo [Tisue and Wilensky, 2004] in contrast uses an own programming language, but provides finished models that can be loaded and explored in a graphical user interface.

3.6 Insights from Related Work

From related work, we can learn that uncertainty communication is relevant for many different application scenarios inside and outside of HCI. However, uncertainty communication and visualization is challenging as, for example, it has many sources, adds an additional dimension to the problem and the visualization, and requires interdisciplinary work. The field of HCI, which only recently started to address uncertainty in interactive systems, can help to tackle some of these core challenges.

Uncertainty can be communicated in various ways with linguistic expressions, numerical probabilities, or complex visualizations. Each form of communication has potential disadvantages, which need to be taken into account when designing a system. In past studies, visualizations such as line charts, dotplots, and probability distribution functions have proved to be promising. One good reason for communicating uncertainty is that users might lose trust in an application if it does not communicate its uncertainty. However, they might also see uncertainty as a shortcoming of an application.

Further studies showed that forecast users are aware of uncertainty and make better decisions if uncertainty information is presented. However, the information should not be overwhelming or confusing. These positive effects of communicating uncertainty indicate that interactive systems should communicate uncertainty to better support humans in decision-making. The previous studies and visualizations can serve as a starting point to explore the communicating of uncertainty in interactive systems.

As this work was carried out in the context of the Cluster of Excellence in Simulation Technology at the University of Stuttgart, we decided to embed our work into the development of an end-user simulation tool. Current simulation tools for end-users are mainly used in education and not suitable for simulating personal application scenarios, but rather help to understand common principles. A more general simulation tool focusing on the communication of uncertainty could serve as a vehicle for future research in this area. In the next chapter, we therefore identify functional and non-functional requirements as another foundation for our work.

Chapter 4

Understanding Simulation Users

In this chapter, we present five connected pieces of research designated to understand simulation users. On one hand, we explored how simulation experts work and what tools they use (see Section 4.2) and on the other hand collected some real usage examples and ideas from the general public (see Section 4.3). We mainly collected qualitative data using questionnaires, a diary study, focus groups, and design workshops.

The main goal of the research described in this chapter is to lay the foundation for the rest of this thesis by understanding prerequisites of simulation usage and potential users. Studying expert users of simulation tools helps to learn about common processes, workflows, pitfalls, and difficulties of using simulation tools. Studying the current usage patterns, understanding, and expectations of the general public helps to understand key requirements and features to lay the ground for developing a simulation tool for end-users.

Parts of this chapter are planned to be published as follows:

- M. Greis, V. Zeamer, N. Henze, and A. Schmidt. Predictive Simulation Services for Everyday Life Usage.

Table 4.1: Detailed research questions for our research on understanding expert simulation users.

Research Question	**Subsection**
What steps of the simulation process do experts work in?	4.2.1
What modeling approaches and tools do experts use?	4.2.1
What is the workflow of simulation experts?	4.2.2
Do experts have and need programming experience?	4.2.2

4.1 Definitions

In this chapter, we distinguish between two potential user groups of simulations: experts and the general public. We define simulation experts as people who regularly create or execute models to run simulations or build simulation tools as part of their work. This includes people working in research and industry. Conversely, we define the general public as people who do not regularly create or use models or simulations in a professional or work-related context, but who mostly have contact with simulations by consulting those created by experts, for example, weather forecasts. We refer to the general public as non-experts.

4.2 Simulation Experts

We conducted an online survey and a follow-up paper questionnaire with simulation experts to gain a better understanding of how they work. Specifically, we were interested in their processes, workflows and requirements for their work. We additionally aimed to understand what tools they use and what knowledge they are required to have for using these tools. The detailed research questions are summarized in Table 4.1. The first inquiry in the form of the online survey consisted of a more general questionnaire to generate first insights. These served as an input for the more specific paper questionnaire to understand relevant details.

4.2.1 Online Survey

To get a first understanding on how simulation experts work, how they create models and how they use simulations tools to run simulations, we created an online survey and shared it with members of the Cluster of Excellence in Simulation Technology at the University of Stuttgart and further simulation experts in research.

Method

We conducted an online survey where we first asked participants for the steps of the simulation process they work on (data gathering, modeling, simulation, visualization, and interpretation) and what exactly they do in these steps. Additionally, we asked them to describe and classify the modeling and simulation method they use during work. We furthermore collected simulation tools that they know and use. At the end of the questionnaire, participants were able to rate up to five tools with the help of a Computer System Usability Questionnaire (CSUQ) [Lewis, 1995].

Participants

Our survey was completed by 48 participants who worked with simulations in a research context. Most of these were PhD students at the University of Stuttgart, but also postdocs, professors and researchers; including some from other universities. Participants had backgrounds in a various amount of disciplines such as Engineering, Natural Sciences, Computer Science, Mathematics, Social Science, and Economics.

Results

We found that 12 participants worked in only one step of the simulation process (mostly simulation), while the rest worked in two or more steps. Most of these worked in either modeling (27 participants) or simulation (33 participants), which was also the most frequent combination (13 participants). Participants specified that their work in these steps includes creating new models, running simulations, improving the simulation infrastructure, or analyzing the impact of simulations.

With the modeling and simulation method, most participants specified that they worked with differential equations. Some of them additionally specified the method that they used for solving these equations, which were mostly finite element method (FEM), finite volume method (FVM), molecular dynamics, atomistic modeling, or Monte Carlo methods.

Participants specified a huge range of tools that they use for creating and running models. Out of 70 mentioned tools, 58 were only named once. The most named tool was Matlab with 13 mentions; all other tools had fewer than five mentions. Additionally, 68.8 % of the participants used self-written programs. The CSUQ questionnaire was filled for 42 tools and the overall average of all questions was positive, which shows that most of the participants were satisfied with their tools. Nevertheless, according to our participants, 57 % of the tools would benefit from improvements such as a better interface (10), advanced features (8), or better documentation (7).

Discussion

We found that simulation experts rarely work in all steps of the simulation process. They often collaborate with other experts who work in the adjunct steps. That also implies that different tools are used for each step. As each step seems to be highly specialized, it is obvious that the general public will not be capable of completing all steps of the simulation process. Especially the modeling step needs expert knowledge and mostly requires programming experience. A possible tool that supports the general public in running simulations clearly has to separate the steps of the simulation process. Other steps such as the modeling step might be still conducted by experts and hidden from non-experts.

Simulation experts also work on improving the infrastructure for their simulations. The general public is not able to improve the infrastructure as they probably do not have programming skills to do so. It is very important that models targeted for the general public need to run on devices and infrastructure that the general public already owns, e.g., smartphones and laptops. Alternatively, web-based tools could make simulation runs on a server where only the results are sent to the user.

With the modeling and simulation method, experts mainly specified to use differential equations, but some also had difficulties to actually specify their method. Differential equations are obviously not suitable for the general public, so ways to reduce the mathematical complexity or hide this complexity from the non-expert user are necessary to make a simulation tool easy to use.

We also found that the tools experts work with are highly specialized and often self-written. They mostly work for one specific problem. For the general public, it would be ideal to have one tool that supports multiple simulations. This would reduce the time to learn the tool as concepts could be transferred every time the tool is used.

4.2.2 Paper Questionnaire

Based on the results of the online survey, we constructed a paper questionnaire to collect more detailed feedback. We gave out the paper questionnaire on a status seminar of the Cluster of Excellence in Simulation Technology from the University of Stuttgart. All participants of the status seminar were simulation experts working with simulations on a mostly daily basis at the university.

Method

The questionnaire that consisted of a two-sided paper. We first asked participants for their current position, their field of work, and what application area of modeling and simulation they were working in. As in the online questionnaire, we asked for the tools that participants use, but this time divided them by the steps in the simulation process. Additionally, we asked them to describe or draw their workflow when working with models and simulations with the help of the tools that they mentioned. We then wanted to know which programming languages they use and how much they agree on a five-point Likert scale with the following statements:

- I am an experienced programmer.
- I need programming experience to complete my tasks.
- I have enough programming experience to complete my tasks.

Participants

In total, 68 participants filled out the paper questionnaire on the status seminar. Most of the participants were PhD students in different fields such as Engineering, Biology, Computer Science, Mathematics, Chemistry, and Physics. They worked on a range of models and simulations with different applications, e.g., in water management, chemical reactions, cancer cell populations, blood flow, or cracks in material.

Results

As already found in our first survey, participants in general indicated that they worked in more than one, but not all steps of the simulation process. Most used different tools for different steps in the process. One participant described this as a tool chain approach where the output of one tool in one step would directly

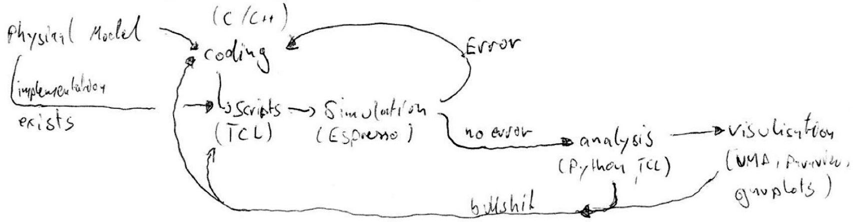


Figure 4.1: Depicted simulation workflow of one participant on the paper questionnaire.

be used as the input of the next tool. However, not all participants used such an approach as not all tools can be used sequentially. Figure 4.1 shows a drawing of one participant that correlates very well to a workflow many participants working in multiple steps of the simulation process described. One important aspect described by another participant was feedback cycles: “Choose appropriate available model, modify model (*), implement model, run simulation, analyze results + visualization, interpret results, repeat from step (*) if needed”. Steps of the simulation process might be repeated if the outcome is not satisfactory.

For the analysis of programming experience, we converted the Likert scale items to numbers: 1 corresponding to “totally disagree” and 5 corresponding to “totally agree”. Participants named 21 different programming languages that they regularly use. The most named languages were C++, Matlab, Python, Fortran, and C. The majority of the participants agreed with the statements that they are an experienced programmer ($M = 3.8, SD = 1.0$), need programing experience to complete their tasks ($M = 4.0, SD = 1.1$), and have enough programming experience to complete their tasks ($M = 4.2, SD = 0.8$).

Discussion

From the workflows provided by participants, we learned that each step of the simulation process has a specific set of tools associated with it. Sometimes tool chains exist that allow for a tool’s outputs to serve as an input for the next tool, but most of the time extra conversion or programing was necessary. We additionally learned that simulation experts use cycles and go back to earlier steps if they realize that they made a mistake in the model or the output is not as expected. Such feedback cycles need to be taken into account and be supported.

Programming experience is widely spread around simulation experts, which cannot be expected from the general public. However, experts’ programming experience is mostly very specialized on the tool that they use and some experts also stated not to have much programming experience. In general, a tool designed for the general public should not rely on extensive programming knowledge as experts might not know the specific language used by the tool. Additionally, this would help domain experts without programming experience to also create models.

4.2.3 Implications from Expert Simulation Usage

Based on the results of the online survey and the paper questionnaire, we developed implications for the future design of simulation tools for the general public by looking at common usage patterns of experts and their difficulties in using simulation tools. In the following, we present the five main requirements.

Adaption for Available Infrastructure. A simulation tool for the general public should run on a device that the general public uses regularly, e.g. a mobile phone or a laptop computer. As non-experts normally do not have the ability to improve the infrastructure of their simulations (besides buying a new device), the models have to be developed in a way to take this into account. Alternatively, web-based tools could compute simulations on servers and only show the results to the users.

Generality. Most simulation tools are specifically built for one use case and one step in the simulation process. For the general public, it would be easier to have a general tool that supports all steps and different contexts. Even experts could profit from such a tool as knowledge transfer would be easier and learning time would be reduced. For non-experts, this knowledge transfer is even more important as they may have more difficulties to understand different tools and cannot create a tool chain to ease the use of multiple tools.

Separate Steps. Non-experts might not have the ability or knowledge to create models. Therefore the model building step should be left to experts. A potential tool needs to carefully separate the steps of the simulation process. Experts could, for example, develop models and then share them with the general public to run their own simulations with the models.

Hide Mathematics. Experts mainly use differential equations to build models. However, the general public usually has no experience in working with differential equations. A method for explaining the models to non-experts has to be found that allows for hiding the mathematics while still getting across the purpose and functionality of the model.

Table 4.2: Detailed research questions for our research on understanding non-expert simulation users.

Research Question	**Subsection**
How does the general public currently use simulations?	4.3.1
What understanding does the general public have of simulations?	4.3.2
What use cases are promising for future simulations?	4.3.2
How do future simulation services have to be designed?	4.3.3

Minimize Programming Knowledge. Simulation experts on average have a high amount of programming experience. A tool designed for non-experts should minimize the amount of programming knowledge needed. This would also benefit experts without programming knowledge and experts with very specialized programming knowledge.

Support Feedback Cycles. When experts work on the simulation process, the work conducted is not completely linear. They often use feedback cycles or iterations to go a step back and start again from an earlier point. This happens if, for example, the model is erroneous or the output is different than expected. A simulation tool needs to support such feedback cycles to be effective.

4.3 Non-Experts

As one of the first steps towards building a simulation tool for the general public, we wanted to understand the current situation and the potential of the usage of simulations in everyday life. We additionally aimed to understand design requirements for simulation services. We therefore conducted a diary study, focus groups, and design workshops focused on predictive simulations. We mainly focused on predictive simulations as these are currently already used by the general public, e.g. in the form of weather forecasts. Our detailed research questions are outlined in Table 4.2. The diary study mostly focused on the first question on how the general public currently uses simulations. The focus groups focused on the understanding of simulations and what use cases could be promising for future simulation services. In the design workshops, we then collected concrete features and prerequisites that are needed for designing such services.

4.3.1 Current Simulation Usage in Everyday Life

We first wanted to get a better understanding of when, why and how the general public already uses simulations in everyday life. We mainly focused on predictive simulations (forecasts) as their occurrences are easy to spot for non-experts. We collected everyday life examples by conducting a diary study and follow-up interviews.

Method

We created DIN A 5 leaflets that included all information about the study, a form for demographic information, four examples of potential diary entries, and 22 empty spaces for participants to note down any occurrences of when they used predictive simulations. In initial sessions with up to six participants, we gave a short oral introduction to the study and defined what predictive simulations are, talked them through the explanation and the examples, and answered questions. Every participant received a leaflet, a pen, and candies as reward for participation.

We asked participants to carry the leaflet with them for one full week to note every occasion in which they used predictive simulations. We called these occasions “information about future occasions” or “forecast” to make it more understandable. To give participants a structure for their diary entries, we asked them to answer the following four questions per entry:

1. At which point in time and in which situation did you need information about the future?
2. Which information did you need?
3. Why did you need this information?
4. How did you access the information?

We used the following examples not to limit the thinking of participants, but help them to understand the task:

- The weather forecast was watched on TV in the evening to know what clothing to wear and whether an umbrella would be needed on the next day.
- The potential driving time from home to a friend’s house was checked on a navigation website to know when to leave home.

- Before going to work, the website of the local public transportation company was consulted for information about train delays to find out if I would be in time for an important meeting.
- Forecasts of the results of soccer world championship were needed after work because colleagues wanted to bet on the results. With the help of a search engine different websites were compared.

All four examples were provided in the way participants were instructed to take their notes. After one week, participants had to return the diary and participate in a short interview. The interview was held in an open form and was used to talk about unclear situations and blanks in the diary. After completing the interview, participants received a mobile phone cleaner as a thank-you gift.

Participants

We recruited 38 participants of whom 25 (18 male, 7 female) handed their diary back to us. Two of the participants who did not hand it back reported that they did not have any occasion to enter in the diary. In the following, we will only report about participants who handed their diaries back to us. The average age of our participants was 31.0 ($SD = 11.8$). Regarding the highest level of education, three of them had completed a secondary school diploma, six a high school diploma, five had received a Bachelor degree, ten had received a Master or equivalent degree and one participant had completed a PhD. Eleven participants were students with different subjects such as Computer Science, Electrical Engineering, Biology, Chemistry, and Management Sciences. The other participants were employees in different fields, including secretaries, teachers, postmen, and logisticians. On average, each participant noted 7.2 ($SD = 4.7$) entries in the diary; between one and 21 entries per person.

Results

In total we received 178 diary entries, but had to exclude 32 from the analysis. The excluded entries did either not cover a situation in which a predictive simulation was used or described occasions in which participants did not receive the information that they needed. In the following, we therefore analyzed 146 diary entries. We coded the diary entries according to the following five categories that we identified from the questions we asked: context, type of information, future point in time, reason, and device.

Context. In 78 entries, participants included a point in time when they needed the information. Besides predictive simulations that were used regularly, specific ones were intensively used during a short period of time. Participant 10 (m, 26 y.), for example looked at the forecasted day of delivery for his mobile phone several times a day until it arrived because he wanted to give his old mobile phone to a friend.

Participants used forecasts after activities in 18 situations. In most of these situations, the forecast was integrated in the daily routine and attached to an activity such as getting up, having breakfast, or arriving at a specific location (e.g. workplace, holiday destination). Participants also referred to very specific situations, e.g. after cutting trees, after a new renter moved into a flat next to them, or after their internet connection stopped working. In these situations, the need of the predictive simulation was triggered by the activity itself. Forty entries contained information about an activity that followed the use of a forecast such as going to a specific location (e.g. workplace, bed, gym), before traveling (e.g. in the car or train), or before having a barbecue. In these occasions, the forecast was mostly used to ensure that future activities could be done or to estimate how long they would take. In 48 entries, participants reported that they used forecasts while, for example, having or preparing a meal; working; planning for weekends, cinema visits, or vacations; sitting in the train or bus; or watching TV.

Type of Information. In total, participants needed 38 distinct types of information during the diary study. In 33 % of the situations, they needed weather information. Other uses included predictive simulations about the delay of public transport (6 %), traffic situation (5 %), and opening hours (5 %), and 12 % of the predictive simulations showed the availability of objects (e.g., appliances, ingredients, free seats in a train) or persons. The majority of the information was only requested once by a single participant, e.g. the development of the gasoline price, the development of the stock price, the expected energy costs, the expected recovery time after being ill, and the potential transfers of football players.

Participant 17 (m, 31 y.) noted one occasion in which he needed more than one type of information at the same time: the weather forecast, the delay of public transport, and a traffic forecast to decide his means of transportation (bike, train, car) for going to work (see Table 4.3). In this specific case, the participant combined multiple predictive simulations to make a more complex decision that could not be solved by a single predictive simulation.

Future Point in Time. 114 entries specified the future point in time that the prediction was made for. Most of these entries corresponded to short-term predictions, more specific in 73 % of the occasions, participants needed information

Table 4.3: Diary entry from participant 17 (m, 31 y.)

Situation	In the morning when getting up.
Information	What's the weather like today? Do the trains run normally? Are there any traffic jams on the way to work?
Reason	Decision what means of transportation is the best to get to the workplace.
Access Method	Weather application on my smartphone and own inspection (sun, clouds, ...), public transportation application, Google Maps, radio traffic service.

Table 4.4: Diary entry from participant 24 (m, 32 y.)

Situation	Lots of work on a Wednesday
Information	How much work do I have to expect during the next weeks?
Reason	To plan leisure time
Access Method	Asking colleagues for their experiences in the last years

about the same day, sometimes only minutes in advance. Only three entries referred to long-term forecasts for a period past the next week, e.g., Participant 24 (m, 32 y.) reported a situation where he needed an estimate of the energy costs for the next winter to calculate house-keeping costs. For some entries, a categorization of the future point in time was not possible as the time itself was forecasts, e.g., the day of delivery of a package, or the recovery time needed when being ill.

Reason. We found five main reasons why participants used predictive simulations: decision-making (e.g., which clothes to wear, which route to take), planning of activities (e.g., schedule appointments, check availabilities), knowledge gain (e.g., when someone returns, potential energy costs), avoidance of unwanted situations (e.g., getting wet, long waiting time), or saving of money and time. Other reasons such as curiosity, boredom, or interest only appeared in one or two entries.

Device. Participants mainly used their computer or their mobile phone (62 situations) to access the forecast information. In 27 situations, participants asked other persons or trusted their own experience for making a forecast, for example participant 24 (m, 32 y.) reported a situation in which he needed information about the workload during the next weeks and asked colleagues for their experiences of the last years (see Table 4.4). In these cases, computer models were not available. Occasionally, participants used written documents, TVs, or radios. In nine situations, participants used more than one device, either because they needed several pieces of information or they did not trust a single source. Other participants also reported using multiple applications on the same device if they did not have enough trust in one source of information.

Interviews

We interviewed 15 out of the 25 participants to discuss their diaries, further forecasts they had already used, and other forecasts they could imagine using in the future. For other forecasts that they had used, most participants mentioned ones reported by other participants in their diaries. Participant 3 (f, 30 y.) additionally mentioned forecasts of the current political development and upcoming elections.

Participants had distinct ideas of what forecasts they would like to use in the future which covered topics such as health, work, finances, etc. Participants also suggested to combining different forecasts, for example weather forecasts, plans of friends, potential costs of activities, and upcoming events to enable their planning for a weekend.

Discussion

In our diary study and the follow-up interviews, we identified two types of users. Half of the participants used forecasts very often, while the other half only used them in extraordinary situations. Based on the reported usage, we therefore identified three usage patterns: *regular usage* (e.g. daily, weekly), *intensive usage during a short period*, or *specific usage* which was highly dependent on the individual, the context, and the activities. Interestingly, participants mostly used short-term forecasts. This could be an indicator that either long-term forecasts do not yet exist or are not used very often. This could be a starting point for offering new services as well as supporting the fact that participants did not necessarily trust a single source or had to combine different forecasts to get to a conclusion. The combination of multiple forecasts could therefore be promising. However, as participants had very distinct ideas for new services, addressing and supporting a larger group of people with one new predictive service could be difficult. Further exploration on potential use cases is necessary.

4.3.2 Developing Definitions and Potential Use Cases

In order to get a better understanding of what people know about simulations and what type of predictive simulations they would like to use in the future, we conducted focus groups.

Method

In total, we conducted three focus groups with up to six participants. Participants were invited by sharing our request on social media and with an e-mail list of prospective participants. The sessions took between 1 and 1.5 hours. After a short introduction, we collected and discussed participants' usage und definitions of predictive simulations. In a phase of idea creation, participants had to imagine predictive simulation services that they would like to use in the future discussing pros and cons of their ideas with a partner.

Participants

Each of the three focus groups was conducted with a different target audience. The first group was held with four German simulation experts (3 male, 1 female) who regularly use simulations at work; the second was held with six students (5 male, 1 female), and the third with four workforce employees (3 male, 1 female) consisting of a postman, a sales woman, and two electrical engineers. We selected these three target groups to get opinions of people with different education levels. All simulation experts had or were pursuing a PhD, the students were completing a college degree, and the employees had some vocational training as their highest education level.

Results

We transcribed the focus groups and collected all material produced by the participants to analyze the data. To extract insights, the data was categorized by two independent researchers who in a second step compared and combined their categorizations. In total, we had 14 participants creating around 60 ideas for new predictive simulation services. In the following, we present their definitions of the term *simulation* and the four most popular ideas for future applications in the form of use cases.

Definitions. Most of the participants had distinct definitions of the term *simulation* and especially the simulation experts did not agree on the definitions of others. Two experts focused on the fact that mathematical functions are used to generate concrete results, while one expert mainly focused on the aspect that a simulation is computer-supported and needs computing power to be conducted. Only one expert used a very formal definition of the term.

> "*A representation of reality/real phenomena in an abstract and reduced manner to analyze properties, structures and mechanisms of*

> *the phenomenon and to make statements about the future behavior of the phenomenon.*" (female, simulation expert)

Participants of the student group either focused on the fact that simulation means to execute a model, is a virtual representation of reality, or makes it possible to observe a process without manipulating the real world.

> "*Test something without consequences in the real world (forecasting)*" (male, student)

Two of the employees mainly focused their definition on the visualization part, representing something virtually or in 3D while the other two focused their definition on the aspect that simulation reproduces functions of a real system. The employees also argued about whether simulations can only happen in the brain when thinking about the outcome of a situation. This reflects a very abstract definition and understanding of the term.

Use Cases. The following four use cases were discussed in at least two of our three focus groups, which highlights their importance.

Use Case 1 - State of Health: Predictive simulations could be used to simulate users' personal health taking into account genetic endowments and their current situations (e.g., nutrition, medicaments). The application would support users in deciding whether to see a doctor or what to do to avoid getting ill.

Use Case 2 - Personal Fitness: Predictive simulations could support sport activities, e.g., by calculating the efficiency of the training beforehand, helping with a healthy execution, or selecting a suitable sport matching personal goals and prerequisites.

Use Case 3 - Personal Finances: Predictive simulations could support financial decisions and financial management, e.g., how long their money might last if a person quits a job or moves to another country, or whether follow-up costs such as insurances and repairs when buying a car will be manageable.

Use Case 4 - Education Path: Predictive simulations could also help to choose an educational or career path, for example how a specific course of studies might influence the later work life.

Discussion

The results of the focus groups revealed that people in general have a broad and very distinct understanding of the term *simulation* independent of their background. One key argument in the discussion was whether simulations have to be computer-supported or can also be performed by thinking. Besides few participants having very detailed or mathematical definitions, they mainly focused on one aspect to define the term *simulation*, for example the virtuality, support by computers, model execution, or 3D visualization. We conclude that the broad term *simulation* should be avoided as it could be misleading, furthermore users of simulation services might also have problems to understand the benefits of a service and to transfer concepts between different services. For example, the term *forecast* seems to be a better choice for predictive simulations. If the general term *simulation* is needed, it should be carefully explained to avoid misunderstandings. As use cases for future development, participants mentioned short-term along with long-term predictive simulations. The most promising application areas are in health, finances, or education-related areas.

4.3.3 Future Usage of Simulations in Everyday Life

Based on the use cases developed in the focus groups (see Subsection 4.3.2), we conducted design workshops to extract a set of critical features and design guidelines for predictive simulation services. We selected the health-related use cases as these were the most prominent ones in the focus groups.

Method

We ran three design workshops with 6 to 12 participants. Each workshop took between 2 and 3 hours. Participants were invited through e-mail and social media channels of the university urging to those interested in predictive health-related services.

All design workshops followed the same pattern. After an introduction of the participants and the topic, participants completed individual brainstorming sessions with one of two design cases that we derived from the focus group results:

1. Find your ideal sport: (a) What is the best sport to lose weight? (b) Should I go swimming, go to the gym, or take a ride with my bike? (c) How often do I have to train to be successful? (d) Does the training suit my body and my individual abilities?

(a) Feature discussion (b) Sketching (c) Preparing for presentation

Figure 4.2: Design workshops about predictive health services for everyday life usage.

2. Avoid illness: (a) What can I do to prevent illness? (b) How can I boost my health?

By introducing different prompts, each participant individually thought about their concerns, relevant items and information, and possible features of an application supporting the design case. Based on their choices, participants were then put into pairs to develop wireframe sketches. At least one participant of each pair had previous sketching experience. At the end of the workshop, each pair presented their wireframe sketches and discussed them with the other participants (see Figure 4.2).

Participants

In total, 26 participants (15 male, 11 female) attended the design workshops. All of them were students, but their courses of studies (e.g., Media Informatics, Software Engineering, Mathematics, Psychology, Education, Industrial Engineering) and education levels differed. Twenty two had prior design experience. The average age of the participants was 22.9 ($SD = 3.3$) years. We compensated them with food and roughly half of the students received class credit.

Results

In total, we collected 13 sets of interface sketches from our participants. Two independent researchers analyzed all sketches by coding the features with a visual and textual coding approach, afterwards comparing and combining the results. In total, we identified eight categories of features included in the designs:

Data Baseline. Each set of sketches included a data baseline feature. We identified four types of input data that were requested for the data baseline in different sets of sketches: data about the user's health (e.g. weight, height, heart rate, etc.), user's preferences (e.g. experience of sports, team or individual sports, etc.), user's emotions, and general constraints (e.g. physical abilities, amount of available money and time, etc.). For entering the data, the interface sketches offered three possibilities: manually by an expert (e.g. doctor), manually by the user, or automatically by connected devices. The discussion of the interfaces revealed that most participants did not want to enter health data manually as it was perceived as cumbersome, but rather would prefer an application to collect the data automatically (e.g. by performing a body scan).

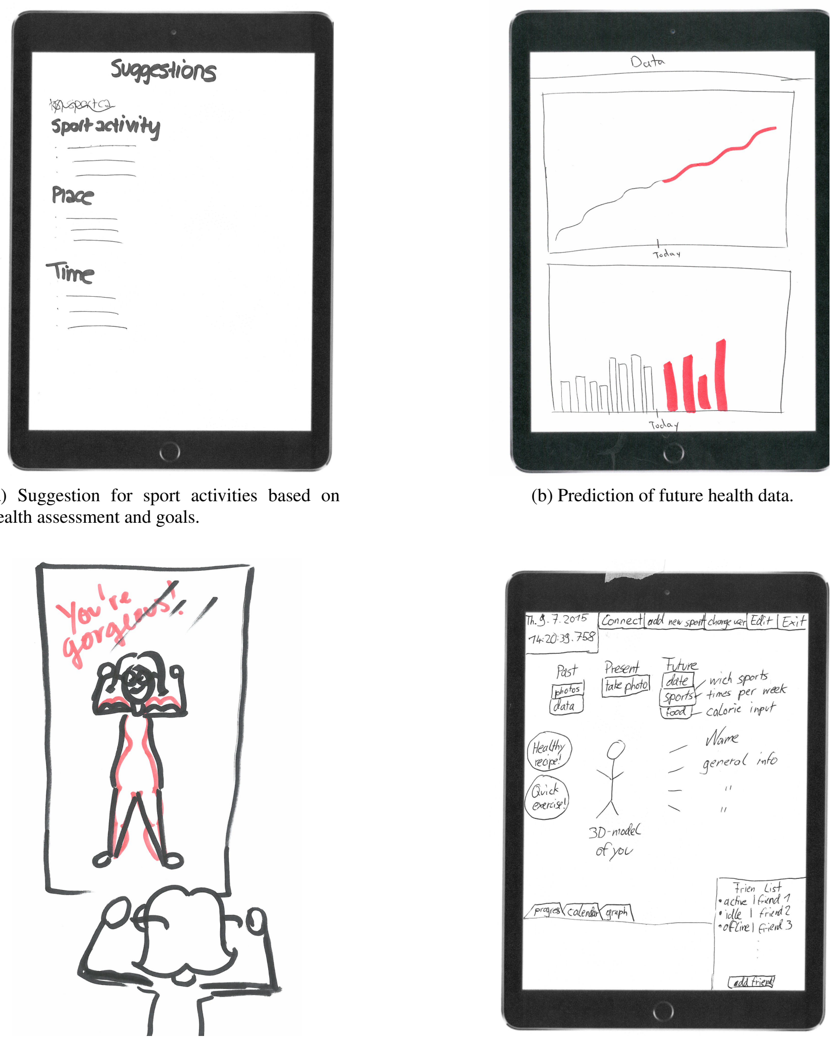
Goal Setting. Seven sets of sketches included a goal setting feature, which mainly consisted of a questionnaire item. Users either could pick a very general goal from a list of goals (e.g. lose weight) or specify a detailed goal (e.g. lose x kilograms).

Tracking. Each set of sketches included a tracking functionality. Mainly the activity, food intake, and environment data were tracked. Most interfaces relied on automatically tracking the data, but three required manual input. In the discussion other participants stated that manual input would be too cumbersome for them.

Prediction. Each set of sketches included a prediction feature. All of the sets of sketches included short-time predictions in the form of suggestions (e.g., a type of activity, healthy food, etc.) as depicted in Figure 4.3a. They mostly focused on the very near future or the next hours. Six interfaces additionally included long-term predictions for at least one week up to one year into the future. Participants envisioned distinct visualizations of the long-term predictions such as charts and diagrams (see Figure 4.3b), a future look of the user's body (see Figure 4.3c), a holistic view of past, current and future self (see Figure 4.3d), or an interactive video. In discussions, participants voiced a preference for future body visualizations instead of charts as they assumed these would boost their motivation.

Execution Plan. Each set of sketches included an execution plan designed to achieve a future goal. Participants mentioned that the execution plan needed to have realistic and believable steps in order to make the predictions trustworthy. The execution plan varied between a full schedule of activities or only activity-based visualizations (e.g. how much did I already cycle this week and how much more do I need to cycle to complete my goal?).

Location. Eight sets of sketches included a feature that made the application location-aware. Participants voiced the concern that without taking the locations



(a) Suggestion for sport activities based on health assessment and goals.

(b) Prediction of future health data.

(c) Visualization of future self in a mirror.

(d) 3D Model of past, current, and future self.

Figure 4.3: Participants' sketches containing predictive features of the interfaces.

into account, predictions could be wrong or unrealistic (e.g. suggestion to go swimming when there is no swimming pool nearby).

Social. Five sets of sketches included a social feature such as a functionality to meet other people, a social media connection, or a chat.

Game. Only one set of sketches included a game, which allowed the user to level an avatar.

Discussion

The findings from our design workshops underline that most potential users wish to have automatic data collection when it comes to parameter input for predictive simulations. Nevertheless, some data such as emotions, personal preferences, or personal constraints, still need manual input which is accepted by users. The predictive aspect of the application has to support short-term and long-term suggestions at the same time and visualize the simulation results in a suitable manner. One of the main challenges of predictive services is to gain the trust of users. The more predictions deviate from the real world and seem unreliable, the more will they lose trust in a prediction. To foster credibility, calculations and results have to be made transparent for users to make sure that they understand the reasoning behind the prediction.

4.3.4 Implications

Based on the results of our diary study, focus groups, and design workshops, we developed implications for the usage and development of future predictive simulation services.

Wording and Understanding

The wording plays an important role for future services. It has to be taken into account that the term *simulation* is a very ambiguous term and can be understood in very different ways. The general public is not necessarily aware of what a simulation is and how it works and might think that it is a 3D visualization. We assume that using more specific words such as *forecast* instead of *predictive simulation* could help non-experts to better understand the premises and dependencies of new services. If the general term is needed as an umbrella term, a very detailed explanation of the term should be given.

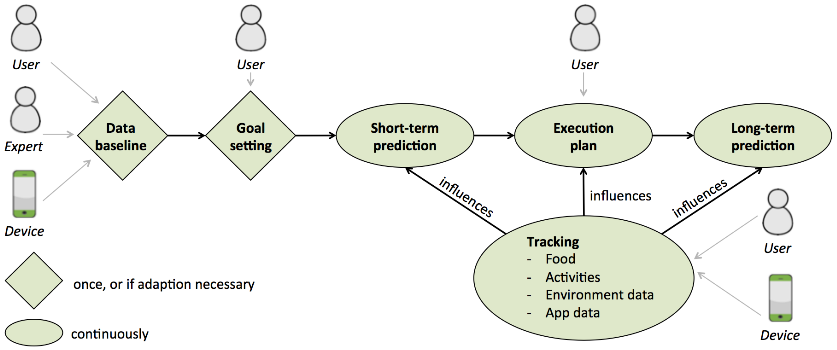


Figure 4.4: Predictive simulation service product flow, highlighting the critical user experience features that need to be offered by a predictive simulation service. The diagram shows the process of interaction with the application. Grey lines show input of either an expert, a user, or a device.

Functional Requirements

Based on the results of the design workshops, we created a set of critical features and a product flow depicted in Figure 4.4. This flow mainly applies for health-related applications, but could possibly be adapted to other application areas. In the following, we outline the four most important aspects.

Sources of Parameter Input. Parameters can be introduced by different sources, either manually or automatically. Most participants in our focus groups stated a clear preference for automatic input as manual input is perceived to be cumbersome. Nevertheless, concurrent information from multiple sources that cannot be tracked automatically still has to be entered manually by the user or a specific expert. The amount of information that has to be entered manually should however be as little as possible or optional.

Illustrative Visualizations. One critical feature is illustrative visualizations. Users prefer visualizations to plain numbers as this makes the results of a simulation more real for them. In the health context, visualizations might also be more motivating for users.

Short-term and Long-term. Although participants mainly already used short-term predictive services, they were interested in using long-term predictive services as well. Complex services may offer both short and long-term services to support the long-term predictions with short-term aims or goals that lead to the long-term prediction.

Personal and Contextual Intelligence. One main feature we identified is the integration of personal and contextual information in predictive services. Especially for short-term predictions, it can be crucial to know the location or environment of a user. Personal and contextual information builds legitimacy for the predictive simulation from the user's perspective.

Non-Functional Requirements

We mainly identified two prerequisites that have to be taken into account when developing predictive simulation services: flexibility and transparency.

Flexibility. First of all, the results from the diary study and the focus groups showed that people already use a whole distinct set of predictive simulations in everyday life and additionally have very individual and specialized needs. A new service has to take into account these individual needs and be able to adapt to different situations and uses. Additionally, it is very important to understand the usage pattern that will be created by a service. Will a service be used regularly as part of a routine, intensively during only a short period of time, or just in very extraordinary situations?

Transparency. Participants in our design workshops stated that a simulation has to be reliable to make them use a service more often. Reliability and increased trust in a service can be achieved by transparency. Making the assumptions of a service and the input data of a simulation transparent for users allows them to better understand how predictions are made and whether they are trustworthy. Uncertainty in the generated results should also not be hidden from users.

Promising Application Areas and Contexts

We identified health, finances, and education as promising application areas. Participants had plenty of distinct ideas in these areas and seemed to be eager to use predictive simulations. We additionally identified two promising contexts.

Forecast Comparison. One promising context could be to allow participants to compare different predictive simulations that use different models. As participants are aware of the uncertainty of predictions, they may not trust a single source and, like in our diary study, use multiple sources. An easy service to compare different predictions may increase their confidence and trust.

Forecast Combination. The second promising context could be to combine different services such as a weather forecast, traffic jam forecast, and prediction of public transportation arrival to support more complex decisions that cannot be solved by a single prediction. Complex services could offer different sources and different predictions for users to make them select their favorite combinations.

4.4 Insights for Developing an End-User Simulation Tool

In this chapter, we analyzed the current usage of simulations by simulation experts and the general public. We conducted several smaller research probes to gain insights into key requirements and promising application areas.

We found that analyzing expert usage provided a deeper understanding on how simulations are currently used and some common pitfalls of simulation tools. To make simulations understandable for the general public, mathematics in the models have to be hidden. We envision a web-based simulation tool that supports different use cases and minimizes programming knowledge by introducing different levels of abstraction.

Non-experts had plenty of ideas on how they could use simulations in their everyday life. We identified the input of parameters and illustrative visualizations as two key aspects that need to be carefully explored. At the same time, participants expected a tool to be flexible and transparent. To achieve these two non-functional requirements in combination with the functional requirements, novel input and output methods are needed. The input and output methods have to be flexible and embrace uncertainty. When uncertainty is taken into account on the input and output level, transparency can be reached much more easily. We therefore identify uncertainty as one of the main challenges in developing an end-user simulation tool.

In the next part of this thesis, we systematically explore the occurrence and the handling of uncertainty in interactive systems. We aim to identify the sources of uncertainty in interactive systems and design and evaluate novel input and output methods that foster transparency on all levels of a system. These methods can be used as a starting point to develop an end-user simulation tool.

III

Uncertainty in Interactive Systems

Chapter 5

Sources of Uncertainty

As related work has shown, uncertainty is an important aspect in different research areas. It has also recently been gaining attention in HCI as more and more interactive systems deal with uncertainty. Psychological research shows that displaying the uncertainty in the results has a positive effect on humans as decision-making improves. Nevertheless, handling uncertainty in interactive systems can only be done if HCI researchers are aware of the sources of uncertainty. In the best case, the uncertainty introduced in different steps has to be quantified and respected when processing data to achieve a reliable outcome. To identify the sources of uncertainty in interactive systems, we build upon the General Interaction Framework introduced by Dix [2009]. In the following, we introduce the General Interaction Framework and, based on the components and translations of the framework, identify potential sources of uncertainty in interactive systems. We further describe more concrete aspects which introduce uncertainty by giving examples and short outlooks on how these uncertainties can become quantifiable.

This chapter is partly based on the following publication:

- M. Greis, H. Schuff, M. Kleiner, N. Henze, and A. Schmidt. Input Controls for Entering Uncertain Data. In *Proceedings of the ACM on Human-Computer Interaction.* 1(1):1–17, jun 2017.

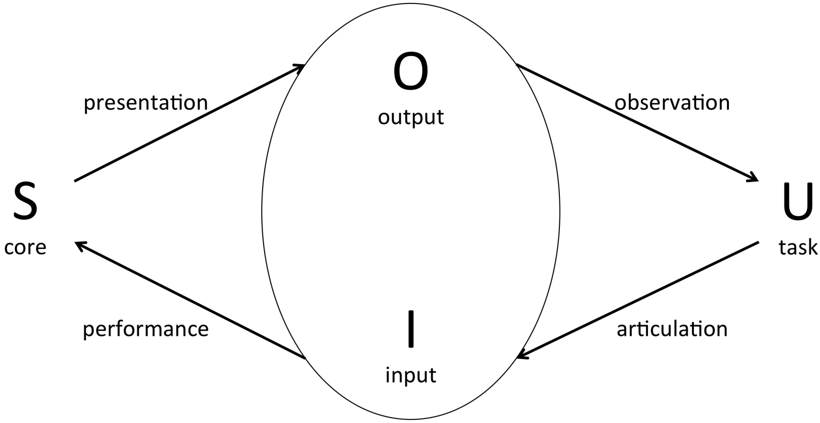


Figure 5.1: The General Interaction Framework by Dix [2009] describes an interactive system by specifying its four major components (system, user, input, output) and the four translations between them (articulation, performance, presentation, observation).

5.1 The General Interaction Framework

Following the definition of the General Interaction Framework by Dix [2009], an interactive system consists of four major components: system, user, input, and output. The input and the output form the interface which separates user and system. The components are connected by four translations:

1. The user has to articulate his/her goals through the input,
2. The system has to interpret the input values and perform an action,
3. The system has to transform the result of the action to the output, and
4. The user observes the output.

5.2 Enhancing the General Interaction Framework

All components and translations in the General Interaction Framework can be affected by uncertainty. Based on related work, we identified how uncertainty plays a role for these components and translations focusing on the concrete sources of uncertainty that need to be quantified or dealt with when developing interactive systems. Finally, we enhanced the General Interaction Framework by adding the sources of uncertainties. This helps to develop interactive systems that are aware and cope with the introduced uncertainty in a way that better suits the user and supports decision-making.

5.2.1 User

One potential source of uncertainty is the user. There are three main user characteristics that contribute to uncertainty: the *educational background*, the *character traits*, and the *culture* of the user.

The *educational background* might impact how the user interacts with a system. Missing domain knowledge, or limited mathematical or statistical knowledge could complicate a user's understanding of the concept of uncertainty and probabilities as whole. Sacha et al. [2016] outline what knowledge could be missing and the influence of the experience level of a user. For example, a user might not understand or be aware that an offered service is based on machine-learning or sensor measurements that introduce uncertainty, and just interact with the system as if it does not contain any uncertainty. Information about the educational background of a user could be retrieved by usage patterns, metadata, and social media accounts. However, calculating its influence might be difficult.

The *character traits* of a user may also introduce uncertainty. Risk tolerance for example influences usage behavior. Users who are less risk-tolerant might not use a system that does not seem trustworthy to them. They might also try different inputs or systems to judge reliability. In general, they probably use the system more carefully or less often. Character traits of a user could also be retrieved by usage patterns, metadata, and social media preferences. Systems could then be adapted to match the character traits of the user.

Besides character traits, the *culture* of a user introduces uncertainty [Bonneau et al., 2014] as different cultures have different preferences and values.

5.2.2 Articulation

During the articulation, sources of uncertainty might be a *lack of knowledge*, *imprecise measurements*, or *limited understanding* which complicate the input of data.

The user might not know the input and therefore a *lack of knowledge* can introduce uncertainty. An example is a library search interface where the user wants to search for an author. A user who does not know how to spell the name of the author might consequently misspell it. Another example would be a user who wants to track calorie intake guessing the weight of the food without measuring it with a scale. The user might be aware or even not be aware of this lack of knowledge. If users are aware of their lack of knowledge, they could communicate this lack of knowledge to the system.

The user might also rely on *imprecise measurements* to articulate the input. An example would be a user calculating calorie expenditure based on the number of steps shown by a step tracker. Step trackers, however, vary in their reliability and might be far off the true value, especially if mobile phone applications are used [Guo et al., 2013]. To quantify the uncertainty, the user could communicate details of the data sources to the system. The system might then be able to determine the reliability of the data source. Boukhelifa and Duke [2009] as well refer to imprecision, for example by using a global navigation satellite system (GNSS) such as the Global Positioning System (GPS), which can act as a source of uncertainty.

Another problem could be a *limited understanding* of the input methods. The user might not correctly understand the required input or the consequences of this input. For example, an interfaces might ask the user to enter a distance in meters, but the user enters the distance in kilometers. The user might also enter a completely different value or make a spelling mistake, which leads to incorrectly entered data [Boukhelifa and Duke, 2009]. A validation of the input could help to detect such obvious mistakes and different strategies could be taken to ask a user to better specify the input if the system is unsure about the understanding of the user.

5.2.3 Input

For the input, uncertainty is mainly introduced due to two factors: *accuracy* and *limited degrees of freedom (DoF)*.

The input could be uncertain due to limited *accuracy*. Sliders and touch screens, for example, only have fixed numbers of pixels and therefore a certain resolution. This could make it difficult to express an accurate input. Schwarz et al. [2010] created a framework for robust and flexible handling of inputs with uncertainty introduced through natural interaction techniques.

The input could also have *limited DoF* [Pang et al., 1997]. A text field, for example, only allows one to enter text while a number field only allows one to enter numbers: they cannot accept both text and numbers. Input fields that allow different values, for example both the deterministic value and the uncertainty of the input, could help to increase the DoF so that a system could quantify the uncertainty in the input.

5.2.4 Performance

For the performance, there are mainly two sources of uncertainty: *transformations* and *imprecise recognition*.

The translation from the input to the system also includes several sources of uncertainty, especially *transformations* [Pang et al., 1997]. An example is a system that expects input in one currency and then transforms it into another currency. The currency conversion rate could be outdated if an old stored version of it is used, or the value could be rounded.

Another source of uncertainty is *imprecise recognition*. When a user is performing gesture or voice input, the system might recognize the wrong input command due to imprecise recognition. Schwarz et al. propose a framework for handling and dealing with such uncertainties [Schwarz et al., 2010]. In cases of user input ambiguity, the system could, for example, give feedback and proactively ask the user to further specify the input.

5.2.5 System

The system itself is also a source of uncertainty. The main two factors are *model uncertainty* and *algorithmic uncertainty* which are always inherently included. Quantifying these uncertainties is a large research field that is gaining attention in different research areas.

The system can include *model uncertainty* [Pang et al., 1997; Wallentin and Car, 2013]. The source of this uncertainty is the real-world model behind the system. Models of the world deliberately focus on several aspects whilst others are left out as not the whole world can be modeled.

Additionally, *algorithmic uncertainty* [Pang et al., 1997] might be added through the concrete implementation of the system. The choice of the algorithm probably has an influence on the calculation (e.g. by using interpolation or extrapolation) as well as the limited precision of computers which could produce overflow or rounding errors.

5.2.6 Presentation

For the presentation the main source of uncertainty is *transformations*, which face the same problems as when transforming the input; the transformations could be outdated or values could be rounded.

5.2.7 Output

The output has the same potential sources of uncertainty as the input as similar problems arise around the two factors *accuracy* and *limited DoF*.

The output could be uncertain due to limited *accuracy* of the output methods such as when the output method can only be displayed in a certain resolution by the output medium [Gershon, 1998]. A bar chart displayed on an LED display with 120 LEDS, for example, will be less fine-granular than on a big 4K monitor.

Additionally, the output could allow for *limited DoF* by providing less DoF that either are calculated by the system or wished for by the user. For example, the output might only allow a standard bar chart without any error bars showing. The system, however, could easily calculate the data needed for error bars internally, and the user would like to see these to get a better impression of the output.

5.2.8 Observation

On the observation phase, uncertainty may arise regarding the *understanding of the output methods* or even the *misjudgement* of the presented information.

The *understanding of the output methods* might be a source of uncertainty if the user does not understand the representation or the presented data. Gershon [1998] names a poor choice of colors or information overload as two examples. Another example is a user looking at a weather forecast interpreting a 50 % chance of rain to mean that it will rain half the day instead of that it rains on 50 % of the days that are similar to the forecasted day [Gigerenzer et al., 2005]. This could be coupled to the educational knowledge of the user. Careful evaluation of output methods and additional explanations on what the output means could help to reduce misunderstanding.

Although the user might understand the presentation, a *misjudgement* of the data is possible as well. The user might misinterpret how high the presented uncertainty is and come to the wrong conclusion. Additionally, users potentially follow their own judgements based on prior experience even if they get the statistically best answer presented as decision advice if they do not know about the uncertainty [Joslyn and LeClerc, 2012]. If, for example, the weather forecast warns that a huge thunderstorm is coming, a listener might not take this too seriously if it is several kilometers away, or the listener might use prior experience of earlier warnings not ending up in a huge thunderstorm to discount this one and follow its advice.

5.3 Implications for HCI Research

Based on the General Interaction Framework, we identified a considerable number of sources of uncertainty in every component of an interactive system. Interactive systems should take these sources into account to better quantify and communicate uncertainty. Figure 5.2 shows the enhanced version of the General Interaction Framework including all sources identified in this chapter. For the field of HCI, we see a need to focus on the articulation and input as well as on the output and observation of uncertainty. This includes developing new techniques for quantifying the uncertainty in user input to increase the degrees of freedom, and developing new techniques for the communication of uncertain data to improve understanding and reduce misjudgments. In both steps, user evaluations are important to verify that new methods are understandable and easy to use. In the following chapters, we therefore present explorations for

1. the articulation and input of uncertain data (see Chapter 6)
2. the communication and output of uncertain data (see Chapter 7), and

3. the interpretation of uncertain data (see Chapter 8).

The explorations add to the body of knowledge on how to design interactive systems that are aware of uncertainty and able to quantify and communicate uncertain data in an easy and usable fashion.

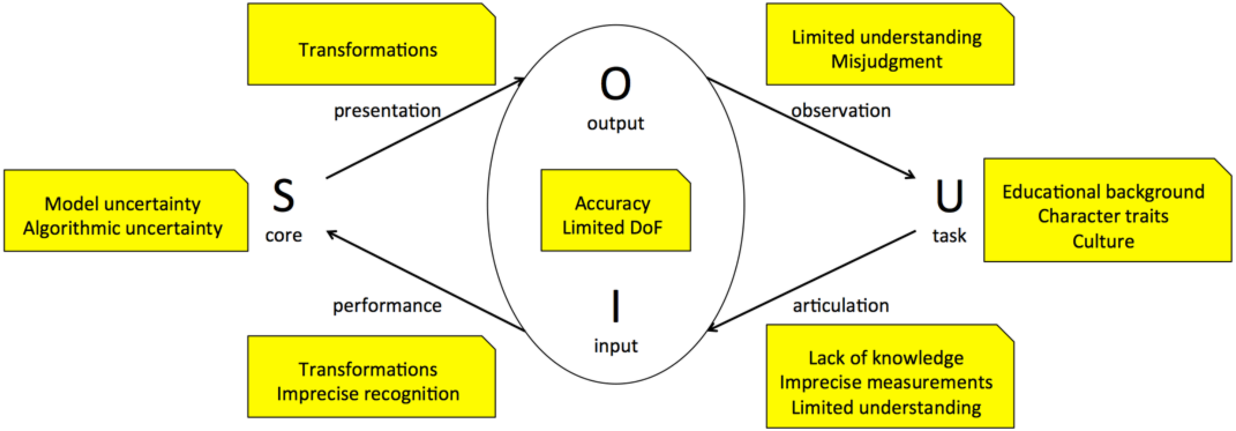


Figure 5.2: The enhanced version of the General Interaction Framework by Dix [2009] including potential sources of uncertainty in interactive systems.

Chapter 6

Input Methods

This chapter describes our exploration of the design space of input methods for uncertain data by conducting four individual research probes. Three probes explore how to actively support users in entering uncertain data. We examine how to enhance common input fields, how to create specialized input controls based on sliders, and whether tangible interfaces are suitable for uncertain input. We additionally present a research probe that examines whether physiological sensing can be used to implicitly measure how uncertain users are. All four research probes include an evaluation in the form of a user study.

The main goal of the research presented in this chapter is to understand how users can be supported in entering uncertainty and whether this could help to make uncertainty in the user input quantifiable. We used different sets of input methods and input modalities to gain an understanding of the feasibility of these approaches. Additionally, we wanted to learn about users' opinions of such input methods.

Parts of this chapter are based on the following publication:

- M. Greis, H. Schuff, M. Kleiner, N. Henze, and A. Schmidt. Input Controls for Entering Uncertain Data. In *Proceedings of the ACM on Human-Computer Interaction*. 1(1):1–17, jun 2017.

Parts of this chapter are also planned to be published as follows:

- M. Greis, H. Schuff, R. Kettner, P. Franczak, and A. Schmidt. Explicit Input of Uncertainty: Enhancing Standard Input Controls.
- M. Greis, J. Karolus, H. Schuff, P. Woźniak, and N. Henze. Detecting Uncertain Input Using Physiological and Behavioral Measurements.
- M. Greis, H. Kim, A. Schmidt, and C. Coutrix. SplitSlider: Exploring Tangible Interfaces for Communicating Input Uncertainty.

6.1 Enhancing Input Controls

Current forms support standard input fields such as text fields and radio buttons. These input controls allow a fixed input of a known answer. However, users may not always be sure of their input; furthermore they may not have the possibility to communicate this to the computer. For example, an interface that tracks users' eating habits provides a form to enter the weight and composition of the food they have eaten. Most users will probably guess the weight of their food, but as a current interface would only offer number fields for the input they cannot communicate the amount of uncertainty included in their input.

The main goal of the work presented in this section is to understand how traditional input fields can be enhanced to offer users the possibility to enter uncertain data. We mainly aimed to understand the preference of potential users for different designs. Therefore, we first of all identified and classified common input fields. Based on the taxonomy, we developed designs and prototypes that we evaluated with potential users by conducting a pre-study and an evaluation in the lab.

In the following, we present our taxonomy of the input fields; our first sketches and the results of the pre-study; the prototype implementation and lab evaluation; and the results of the in-the-wild evaluation.

6.1.1 Common Input Controls

To enhance existing input controls, we first looked at different developer guides to identify common input controls. We then developed a taxonomy for these input controls based on the input they accept. For each cluster of the taxonomy, we developed prototypes for entering uncertain data.

Identification of Input Controls

To identify common input fields, we looked at the documentations and developer guides of Windows[1], HTML[2], Android[3], and iOS[4]. Table 6.1.1 shows an overview of the input controls that are commonly used. All of the development guides, e.g., contain buttons such as action buttons, radio buttons, and checkboxes. Less common are specific controls such as list boxes or ratings. In the following, we shortly describe the input controls.

Textfield. A general textfield allows users to input any kind of string. However, specific versions of these textfields exist. Password fields, for example, do not show the input in the field but instead show placeholders and e-mail fields only allow the user to enter valid e-mail addresses, which corresponds to a string with a specific pattern. Textfields could either be one line or span multiple lines.

Button. An action button is a button that activates a specific function if clicked (e.g. save an entry, undo, etc.). A toggle/switch button allows toggling between two states (e.g. on and off). Radio buttons instead allow users to change between different states. They are summarized in groups and only one of them can be selected per group . Normally they are depicted as a circle. Checkboxes are quite similar to radio buttons, but they do not need to be grouped and even if they are

[1] Windows developer guidelines:
`https://msdn.microsoft.com/en-us/library/windows/desktop/dn742399(v=vs.85).aspx`

[2] HTML form elements: `https://www.w3schools.com/html/html_form_elements.asp`, HTML input elements: `https://www.w3schools.com/html/html_form_input_types.asp`

[3] Android developer API guides: `https://developer.android.com/guide/topics/ui/controls.html`, `https://developer.android.com/reference/android/widget/package-summary.html`

[4] iOS human interface guidelines:
`https://developer.apple.com/ios/human-interface-guidelines/ui-controls/buttons/`

Table 6.1: Overview of input controls explained in developer guidelines.

		Windows	**HTML**	**Android**	**iOS**
Textfield	*General*	✓	✓	✓	✓
	Password		✓		
	Email		✓		
	Search	✓	✓		
	Telephone Number		✓		
	URL		✓		
Button	*Action Button*	✓	✓	✓	✓
	Toggle/Switch Button			✓	✓
	Radio Button	✓	✓	✓	✓
	Check Box	✓	✓	✓	✓
Number Picker	*Stepper/Spin Control*	✓	✓	✓	✓
Slider	-	✓	✓		✓
Dropdown	-	✓	✓		
Picker	*Date*		✓	✓	✓
	Time		✓	✓	✓
	Color		✓		
	File		✓		
List Box	-	✓		✓	
Rating	-			✓	

grouped, multiple checkboxes can be selected. A single checkbox could work as a toggle button. Checkboxes are normally depicted as squares.

Number Picker. A number picker, also called stepper or spin control, allows users to only enter numeric values. Sometimes these values could be restricted to a certain range.

Slider. A slider allows a user to pick a point on a continuous scale. This could either be a numeric value or any other continuous scale (e.g. light to dark).

Dropdown. A dropdown is a folded list. The user can open the list and select one of the options to be shown in a kind of text field belonging to the dropdown.

Picker. Different UIs support pickers; such as a date picker, color picker, or file picker. In general, these could as well be seen as dropdowns with more options or nicely formatted lists to select from.

List Box. The Microsoft guidelines also contained a list box. The list box allows a user to pick one or multiple items of a list. In contrast to a dropdown, the list is not folded but always visible.

Rating The Android developer guide contained a rating as an input control. This could, for example be a star rating with five stars.

To enhance these input elements to support users at entering uncertainty, we decided to first build a taxonomy of these input controls based on what input they allow. Controls accepting the same input could be combined with the same methods for entering uncertainty.

Taxonomy of Input Controls

Depending on the possible input options that the input controls allow, we divided the input controls into four clusters. To build the clusters, three researchers sorted the input controls according to their possible input in a joint session. The sorting continued until the agreement of all researchers was reached. The clusters are not necessarily exclusive we however added input controls to the cluster which they fit best.

Cluster 1: Selection. The user can select m pre-defined elements of the finite set $M = \{E_1, E_2, \ldots, E_n\}$ with $n \geq 1$. The user is able to select the subset $M_s \subseteq M$. We divided this cluster into two subclusters: *Subcluster a)* only contains input controls that allow the user to exactly select one element ($|M_s| = 1$); *Subcluster b)* contains input controls that allow the user to select at least 1, but also up to $|M| = n$ elements ($1 \leq |M_t| \leq n$).

a) Selection of one element: $|M_s| = 1$
Examples: Radio buttons, Dropdown, Listbox (single selection), Color Picker, File Picker

b) Selection of one or multiple elements: $1 \leq |M_s| \leq n$
Examples: Check boxes, Listbox (multi selection)

Cluster 2: Interval Input. The user can pick a value from the interval $[a,b]$. We divided this cluster into two subclusters: *Subcluster a)* only contains input controls that support to select a finite interval $[a,b], a < \infty, b < \infty$; *Subcluster b)* contains input controls that theoretically allow to enter an infinite interval $[a,b], a \leq \infty, b \leq \infty$. Practically this interval is not infinite due to a definition of a lower and upper bound or technical constraints.

a) Finite interval: $[a,b], a < \infty, b < \infty$
Examples: Slider, Rating, Time Picker

b) Infinite interval: $[a,b], a \leq \infty, b \leq \infty$
Examples: Date Picker, Number Picker

Cluster 3: Character Input. The user enters a sequence of characters. We divide this cluster into two subclusters: *Subcluster a)* contains input controls that in general allow to enter an arbitrary sequence of characters ($.^{*}$ or $.^{+}$); *Subcluster b)* contains input controls that restrict the sequence of characters by applying regular expressions, such as $[a-z]\{5,14\}$.

a) No restrictions: $.^{*}$ or $.^{+}$
Examples: General Textfield, Search Textfield

b) Pattern restrictions: e.g. $[a-z]\{5,14\}$
Examples: Password Textfield, Email Textfield, Telephone Number Textfield, URL Textfield

Cluster 4: Action Trigger. The user triggers an action, which can be a primary or a secondary action. We therefore divide this cluster into two subclusters: *Subcluster a)* contains all input controls that allow to trigger a primary action, e.g., “Submit”; *Subcluster b)* contains all input controls that trigger a secondary action, e.g. toggling a state between “ON” and “OFF”.

a) Trigger a primary action: e.g. “Submit”
Example: Action Button

b) Trigger a secondary action: e.g. Toggle state between “ON” and “OFF”
Example: Toggle/Switch Button

As cluster 4 does not really correspond to input, but rather action items, we decided to exclusively focus on the first three clusters. In the following, we will therefore refer to methods that apply for cluster 1, cluster 2, and cluster 3.

6.1.2 Methods for Entering Uncertainty

We focused on three different methods on how input uncertainty could be entered by users: entering a textual description that explains the uncertainty, entering additional numeric values, or by allowing multiple answers. We developed these methods based on informal communication with potential users, and brainstorming sessions. We decided to only focus on the explicit input of uncertainty as functionality such as validation or autofill cannot substitute explicit methods.

Text

The uncertainty can be specified by allowing the user to enter a textual description of the uncertainty.

1) Textual Description. In addition to the input, the user can enter an optional text on why and how the input is uncertain. An example could be an interface for fitness tracking where the user can enter a statement about the uncertainty of the input data, e.g., the user might not have used a scale to weigh the food and therefore guesses the number of grams.

Numerical Values

The uncertainty of the input can be specified through the user by adding one or multiple numerical values to the input. The additional single value can, e.g., be a probability percentage. Multiple numerical values could be used to enter a range instead of a single value.

2) Single Value. In addition to the input, the user specifies a single numerical value which represents the uncertainty. This could be, for example, the probability percentage of the uncertainty, in which 0 % represents a completely certain answer; 100 % a very uncertain answer.

3) Range. Instead of a single value, the user could also enter a range or interval for the input; the bigger the interval, the bigger the uncertainty. The interval bounds would correspond to the minimum and maximum values that a user expects for the input; the maximum interval should correspond to the values that the input can accept.

Allow Multiple Answers

The uncertainty of the input can be specified by allowing users to give not only a single answer, but multiple possible answers.

4) List. Instead of a single input, the user could enter multiple values. For example, if the user is unsure whether a name is spelled with an “e” or “i”, the user might just enter both names instead of deciding which one to enter.

5) Ranking. Instead of a single input or a list of possible values, the user could do a ranking of the input based on probability. In the former example, the user might be more sure that the name contains an “e”.

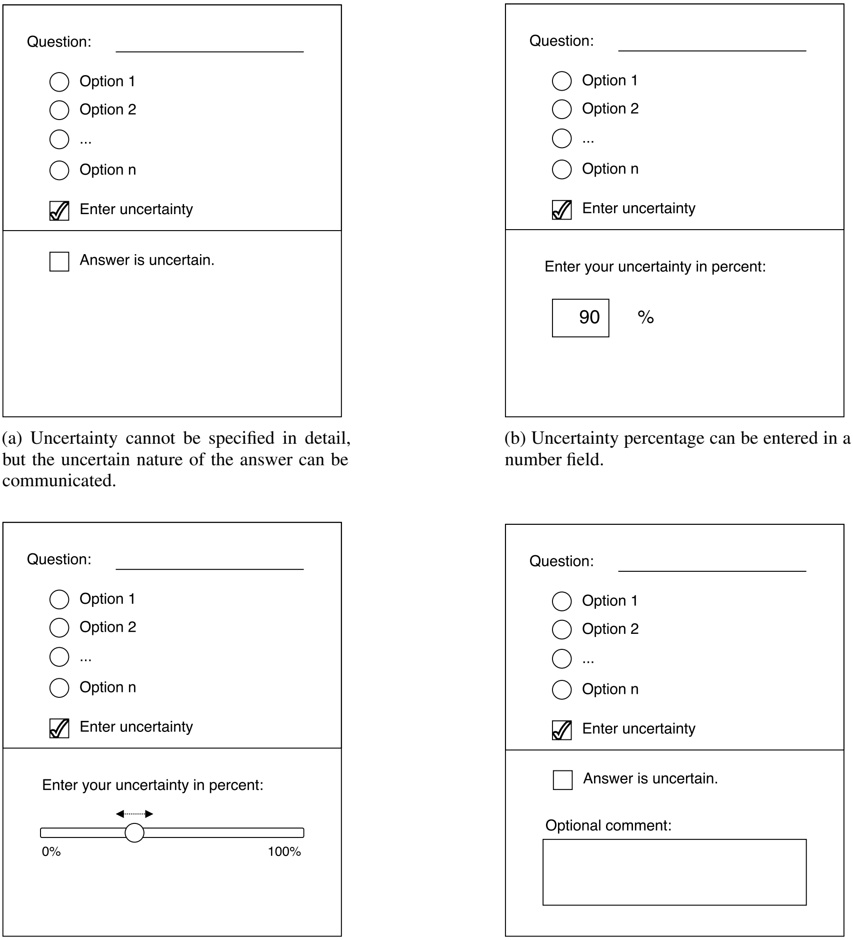
Table 6.2: Suitability of uncertainty input methods 1) to 5) for the identified clusters of input methods: 1) textual description, 2) single value, 3) range, 4) list, 5) ranking.

		1)	2)	3)	4)	5)
Cluster 1: Selection	*a)*	✓	✓			✓
	b)	✓	✓			✓
Cluster 2: Interval Input	*a)*	✓	✓	✓		
	b)	✓	✓	✓		
Cluster 3: Character Input	*a)*	✓	✓		✓	✓
	b)	✓	✓		✓	✓

6.1.3 Design of Non-Functional Prototypes

Not every input control can be used in combination with every method for entering uncertain data. As depicted in Table 6.2, a textual description and a single value could be used with each of the input controls. Range selection, however, only makes sense for interval data. Lists and rankings mostly correspond to selection tasks and textual input. For these respective combinations, we developed non-functional digital prototypes. However, we decided to further limit the scope by only developing prototypes for the selection and the input interval cluster. Character input is a good candidate for validation and autofill or regular expressions related to the uncertainty input methods of lists, however, our developed methods are not particularly suitable for character input.

We developed a set of non-functional prototypes consisting of 32 sketches. Figure 6.1 shows a subset of the developed sketches using radio buttons as the main input control. Figure 6.1a shows a sketch where a user could only state that the answer is uncertain, but not give details about the uncertainty. We used this as a baseline. Figure 6.1b and Figure 6.1c show an interface that allows the user to enter a percentage value with a number field or a slider. Figure 6.1d includes a text field where a user could leave an optional comment about the uncertainty. We designed similar paper prototypes for the other clusters for the following six input controls: radio buttons, drop-down list, check boxes, list box, slider, and number picker.



(a) Uncertainty cannot be specified in detail, but the uncertain nature of the answer can be communicated.

(b) Uncertainty percentage can be entered in a number field.

(c) Uncertainty percentage can be entered with a slider.

(d) Information about uncertainty can be included in a text field.

Figure 6.1: Examples of the non-functional prototypes for the discussions with potential users. The prototypes all use radio buttons and an additional uncertainty input method.

6.1.4 Selection of Promising Designs

To identify the most promising designs from our prototypes and conduct a lab study, we conducted a pre-study in which we discussed all sketches with potential users.

Method

We first asked participants for demographic information such as age, gender, and background. We then focused on a few introductory questions before giving participants an overview of the clusters and the idea of uncertain input. For all groups of paper prototypes (with the same method of uncertainty input), we asked the participants to describe their first impression, their reasons for liking or disliking the prototype, and whether they would use it. Additionally, participants were to indicate how much they liked the presented interfaces and how easy they tought it would be to use on two five-point Likert items.

Participants

In total, we interviewed eight participants (6 male, 2 female). On average, their age was 21.6 ($SD = 2.0$) years. All were students of the University of Stuttgart and had a good general knowledge about user interfaces and input controls.

Results & Discussion

In the discussions, participants expressed that they mostly liked user interfaces developed for cluster 1a) and cluster 2a) (see Table 6.3 for the results). Especially well perceived were the probability percentage input and the range input. Participants did not like the simple check box to indicate uncertainty. They also criticized more complex interfaces developed for the clusters 1b) and 2b). We therefore decided to focus on the interfaces for clusters 1a) and 2a), which were the most favored by the participants. More theoretical groundwork may be necessary to find suitable methods for entering uncertainty for the other clusters.

6.1.5 Evaluation in the Lab

As a second evaluation, we conducted a user study in the lab selecting the most promising paper prototypes from the interviews. We implemented functional versions of these paper prototypes in a web-based study interface, which the

Table 6.3: Results from the pre-study on how much participants liked a specific combination of main input method for clusters (top) and uncertainty input method (left). Likert items were converted to numbers from 1 to 5; 1 corresponds to "not at all" and 5 to "very much".

Uncertainty Input Method	Control	1a)	1b)	2a)	2b)
Baseline	*Check box*	2.25	2.13	2.38	2.38
Textual Description	*Text field*	3.13	2.88	2.63	2.63
Ranking	*Ranking list*	2.63	2.63	-	-
	Percentage sliders	3.25	3.25	-	-
Probability percentage	*Number field*	3.88	3.00	-	-
	Slider	3.88	3.75	-	-
Range	*Pair of number fields*	-	-	3.5	3.5
	Two-thumb slider	-	-	4.38	4.38

participants then used to answer questions while we collected quantitative and qualitative data about their usage.

Method

We conducted a user study with in total 60 questions out of the fields of sports and nutrition. For each question, participants had to enter the answer and their uncertainty about the correctness of their answer. The study consisted of two different parts.

In the first part, participants had to answer 30 multiple-choice questions. Half used radio buttons; the other half used a drop down menu. They had to select the correct answer out of four possibilities. For all questions, participants had to specify their uncertainty as a percentage value between 0 %-100 %: 0 % for a completely certain answer; 100 % for a completely uncertain answer. For reporting their uncertainty, participants used two different user interfaces for 15 questions each: a numeric slider, and a number field. Figure A.1a shows the numeric slider and Figure A.1b shows the number field.

In the second part of the study, participants had to answer 30 questions with numeric answers. Half used a numerical sliders; the other half used a number field. For all questions, they had to specify their uncertainty as a range of numerical values in addition to the point estimate. For the question "How much magnesium (in mg) does a 100 g mango contain?" they could provide a point estimate in the range between 0 mg-100 mg (e.g. 20 mg) and then specify an additional range (e.g. 5 mg- 30 mg) to describe their uncertainty. For reporting their uncertainty, participants used two different user interfaces for 15 questions each: a numeric slider with two thumbs, and a pair of number fields. Figure A.2a shows the numeric slider with two thumbs and Figure A.2b shows the two number fields.

We used a Latin square pattern to equally distribute the assignment of the main input methods and the ordering of the additional uncertainty input methods to participants. The order of the questions was randomized per part.

Besides demographic data, we collected specific data per question such as the used interface components, the answer, the reported uncertainty, and the time needed to answer the question. Additionally, participants completed an Usability Metric for User Experience (UMUX) questionnaire [Finstad, 2010] after each 15 questions, which corresponded to using one type of additional input control for reporting uncertainty.

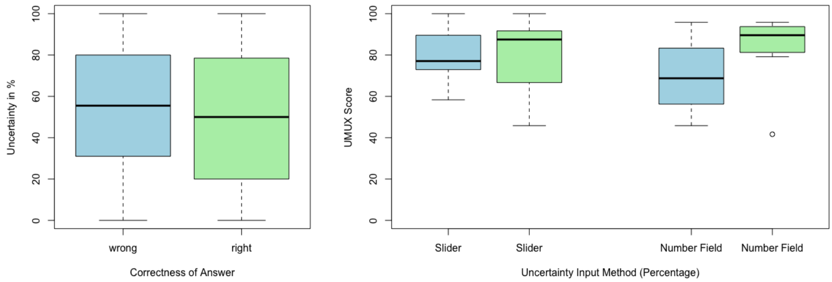
Participants

In total, we conducted the user study with 16 participants (6 female, 10 male) with an average age of 22.9 ($SD = 2.8$). Participants were recruited in a university setting, so the majority of them were undergraduate students.

Results

We used R and lme4 [Bates et al., 2015] to perform a linear mixed effects analysis. In all models, we added subject ID and question ID as random effects and the respective variables as fixed effects. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. For the UMUX score, we performed linear mixed-effects model analyses on the aligned-rank transformed data [Wobbrock et al., 2011].

Multiple-Choice Questions. For the multiple-choice questions, we did not find any significant effect of the main input method and the uncertainty input method on the uncertainty percentage. People entered quite similar values for the percentage slider ($M = 52.6\,\%$, $SD = 31.7\,\%$) and the number field ($M = 53.6\,\%$, $SD = 31.1\,\%$). We additionally did not find a significant effect of the main input method and the uncertainty input method on the time that people needed to enter their answer. For the slider, people on average needed 22.2 s ($SD = 20.0s$) and for the number field 20.9 s ($SD = 13.4s$).



(a) Provided uncertainty percentage grouped by whether participants' answers were right or wrong.

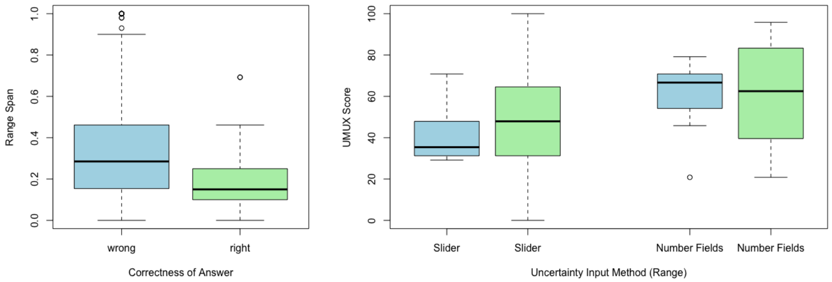
(b) UMUX scores grouped by uncertainty input method and primary input method (blue - drop down menu, green - radio buttons)

Figure 6.2: Results for the first part of the experiment where participants had to select one value and used a slider or a number field to enter a percentage value.

In 172 trials the multiple-choice question was answered correctly, and answered wrong in 308 trials. The mean uncertainty percentage differed significantly for correct and wrong answers according to a Welch's two sample t-test, $t(336.78) = -3.115$, $p < .01$ (see Figure 6.2a). For correct answers, the entered uncertainty percentage was significantly smaller ($M = 47.1\,\%$, $SD = 32.3\,\%$) than for wrong answers ($M = 56.5\,\%$, $SD = 30.4\,\%$).

The UMUX analysis revealed no significant main effects or interaction effects for the UMUX score (see Figure 6.2b). The percentage slider reached a UMUX score of 79.7 ($SD = 13.1$) in combination with the drop down menu and the same UMUX score in combination with the radio buttons ($M = 79.7$, $SD = 18.7$). The number field received a lower score in combination with the drop down menu ($M = 69.8$, $SD = 17.8$) than with the radio buttons ($M = 83.3$, $SD = 17.8$).

Numerical Questions. For the numerical questions, we did find a significant effect of the uncertainty input method on the size of the range the subjects entered, $p < .05$. We normalized the range span to a value between 0 and 1 where 0 corresponds to the lower bound and 1 to the upper bound of the range. People entered a significantly larger range with the two-thumb slider ($M = 0.4$, $SD = 0.3$) than with the pair of number fields ($M = 0.3$, $SD = 0.3$). We did not find a significant effect of the main input method and the uncertainty input method on the time that people needed to enter their answer. For the two-thumb slider, people on average needed 27.2 s ($SD = 21.1\,s$) and for the number fields 28.7 s ($SD = 32.0\,s$). Nevertheless, the difference between the bounds of the range and



(a) Provided range span grouped by whether participants' answers were right or wrong.

(b) UMUX scores grouped by uncertainty input method and primary input method (blue - numerical slider, green - number field)

Figure 6.3: Results for the second part of the experiment where participants had to enter a numerical value and used number fields and a two-thumb slider to enter a range for their uncertainty.

the correct answer if the answer was not in the range was similar for both controls ($M = 0.1$, $SD = 0.2$).

In 30 trials the numerical question was answered correctly, and answered wrong in 450 trials. In 197 of the wrong trials, the answer was in the selected range. The range span differs significantly for correct and wrong answers according to a Welch's two sample t-test, $t(38.51) = -4.337$, $p < .001$ (see Figure 6.3a). For correct answers ($M = 0.2$, $SD = 0.2$), the entered range span was significantly smaller than for wrong answers ($M = 0.3$, $SD = 0.3$).

The UMUX analysis revealed a significant main effect for the method of uncertainty input on the UMUX score, $F_1 = 5.71$, $p < .05$ (see Figure 6.3b). Participants rated the number fields significantly higher ($M = 60.7$, $SD = 5.7$) than the two-thumb slider ($M = 44.8$, $SD = 5.9$). We did not find a significant interaction with the primary input method.

Discussion

The results indicate that for the percentage input, both input methods work well. They do not differ significantly in the entered percentage, input time, or UMUX. Both methods received on average a good UMUX score (according to the System Usability Scale (SUS) values by Bangor et al. [2008]). The primary input method seems to be not tightly connected to the uncertainty input, however the results indicate that there could be a difference (at least for the number field), which should be evaluated with further studies.

For the second part, the results differ slightly from the first part. Again, there is no significant difference in the input time of the two uncertainty input methods, however, they differ in the selected range span. We assume that the slider is not as precise as the number field and that people are biased to enter even values into the number field while they do not care that much about uneven numbers when dragging the slider. Both controls therefore could be helpful depending on whether precision or debiasing is the aim of an interface. However, the slider overall received a poor UMUX rating, while the number fields still reached an acceptable value. The slider might have been too complicated and decoupled from the task. To improve the interfaces, we suggest to develop interfaces that allow the input of the deterministic value and the uncertainty value at the same time. The acceptance of the controls might also change if users are not forced to enter uncertainty but can decide on their own whether to enter uncertainty or not. The additional controls in this case make the user interface more complex and difficult to understand.

6.1.6 Implications

This part of work showed that different options exist to enhance common input fields of different categories to allow for uncertain input. However, people rather prefer quantitative methods for entering uncertainty, as qualitative feedback requires a huge effort. Additionally, people preferred to have input methods for uncertainty that they are already familiar with such as slider and number fields. More complex prototypes did not receive good feedback. We therefore concluded that it is easier to develop uncertainty input methods for selection tasks or numerical input.

The percentage input for uncertainty was well perceived and both developed user interfaces received a good UMUX score in the study. This input can also be easily used by systems to calculate results with uncertainty and definitely has potential to be used in future interfaces. By providing a default selection of 0 %, users would know that they do not even have to bother interacting with the additional control but could use the standard control only.

Although our range input methods did not score very well in the UMUX, we argue that range input could be helpful in the future. We assume that uncertainty input methods allowing range input need to be more connected to the actual value input to not require too much additional input. This could probably be achieved by designing new input controls specifically made for uncertain input.

6.2 Probability Distribution Sliders

We took a second step to explore input controls for uncertain data. In contrast to the controls presented in Section 6.1, we used a mathematical foundation and concentrated on feedback and transparency of how the input would be interpreted by a system. Instead of comparing completely different designs, we decided to focus on one input control and compare different degrees of freedom.

Sliders are flexible input controls that are currently used for a variety of tasks. Several of their properties make them an ideal input control for uncertain data achieving transparency. Sliders are mostly known as visual analogue scales, used in clinical trials and research [Marsh-Richard et al., 2009]. Their usage has been studied widely, e.g., revealing that tick marks introduce a bias [Matejka et al., 2016]. But sliders are also known as selection controls for data exploration, for example the alphaslider [Osada et al., 1993] allows a user to select words, phrases, or names from textual lists. Sliders can also display the distribution of existing data to filter by showing a density plot in the slider bar [Eick, 1994]. A similar approach is suggested by Willett et al. [2007]. Their scented widgets incorporate visual elements that support the selection and exploration of data. Lasram et al. [2012] also enhanced the slider bar with additional visualizations to show the effect of the input on an image. These combinations of sliders with visualizations are especially promising for the input of uncertain data.

We therefore decided to use sliders as the basis for our work and develop slider controls that offer different degrees of freedom. The main goal of this part of work was to understand how users interact with enhanced slider controls that allow manipulation of properties of a probability distribution function and how they handle different degrees of freedom. To evaluate our designs, we conducted an online survey and a controlled user study in the lab.

6.2.1 Design Process

Common input controls for numerical input allow users to enter a single value, which could either be the mode, the median, or the mean (expected value) for

Table 6.4: Deriving levels with varying degrees of freedom for entering a probability distribution function.

	SD	Skew	Kurtosis
Level 0	not included	not included	not included
Level 1	fixed	fixed	fixed
Level 2	adjustable	fixed	fixed
Level 3	adjustable	adjustable	fixed
Level 4	adjustable	adjustable	adjustable

probabilistic input. We based our designs on different degrees of freedom that can be derived from the properties of a probability distribution function: mode, standard deviation, skew, and kurtosis. Each of these properties adds additional flexibility and could be either not included, fixed, or adjustable by the user. In total, we derived five levels with rising flexibility and complexity listed in Table 6.4.

As a baseline for our designs, we used a standard web slider. Based on the four levels, we developed four additional input controls (ICs). The number of an IC corresponds to the levels of Table 6.4. The standard web slider corresponds to IC0.

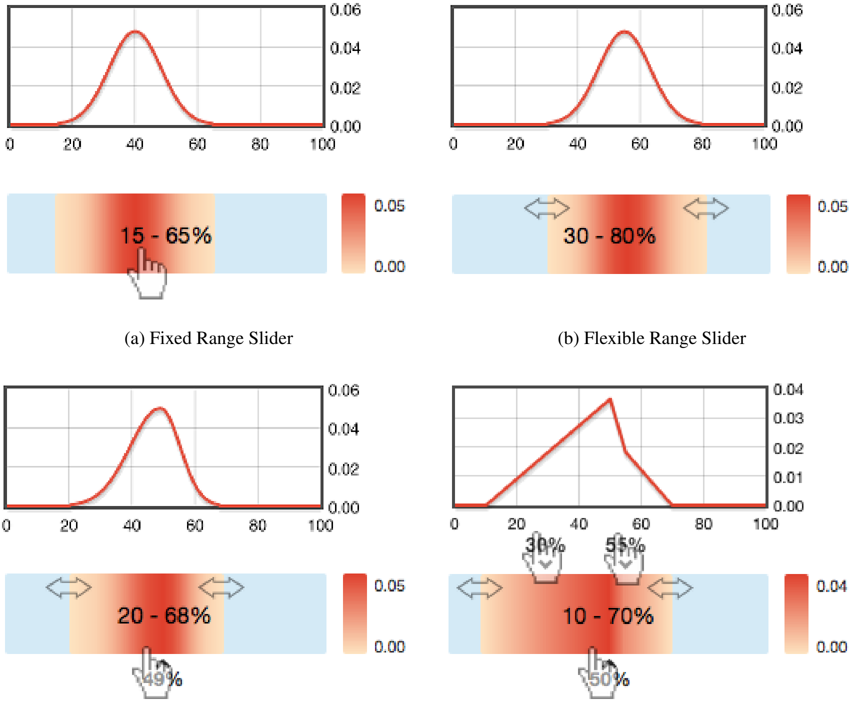
IC1 (see Figure 6.4a) allows the user to move a fixed selection of the slider bar. Standard deviation, skew, and kurtosis of the probability distribution function are fixed.

IC2 (see Figure 6.4b) allows the user to drag at the two ends of the selection to create a range with flexible size. This influences the standard deviation of the probability distribution function.

IC3 (see Figure 6.4c) allows the same interaction as IC2, but additionally offers to specify the mode. By specifying the mode, users can influence the skew of the distribution to create asymmetric distributions.

IC4 (see Figure 6.4d) allows the same interaction as IC3, but additionally offers two more values of the distribution (half as high as the mode), which can be specified. This influences the kurtosis of the probability distribution function.

To achieve transparency of how the input will be interpreted by the system, we added the same three supportive visual elements to every IC: (1) A gradient plot,



(a) Fixed Range Slider

(b) Flexible Range Slider

(c) Flexible Range Best Estimate Slider

(d) Advanced Flexible Range Best Estimate Slider

Figure 6.4: Four slider controls each enabling users to enter a probability distribution function. The sliders have ascending degrees of freedom (left to right). The interaction cues shown in the pictures are displayed when hovering over an interactive element (drag handle) of the slider control.

(2) a gradient height legend for the gradient plot, and (3) a plot of a probability distribution function. In addition to these visual elements supporting transparency, we added interaction cues and tooltips. Before a user interacts with a control, the tooltips will be displayed to help a user understand the control. During the interaction, the interaction cues are displayed when hovering over the interactive elements of the IC.

6.2.2 Online Evaluation

As a first evaluation for the slider controls, we conducted an online survey, in which we collected subjective feedback according to perceived effectiveness, efficiency, ease of use, satisfactions, and learnability of the controls.

Method

We first asked participants of the online survey to provide demographic information and assess their knowledge about stochastic, statistics, probability theory, and probability distributions. We then presented the sliders (IC0 to IC4) in randomized order to reduce sequence and learning effects. Each page contained a short description of the input controls and an exemplary task. Based on a table that showed how often “Sam Sample” used his car over 36 months, we asked the following question: “How many times each month does Sam Sample use his car to go to work?”. The participants were encouraged to try out the control and indicate their level of agreement on the following five-point Likert type items:

Effectiveness: “I am confident that I am able to correctly enter data with this input method.”

Efficiency: “I was able to quickly enter data using this input method.”

Ease of use: “It was simple to use this input method.”

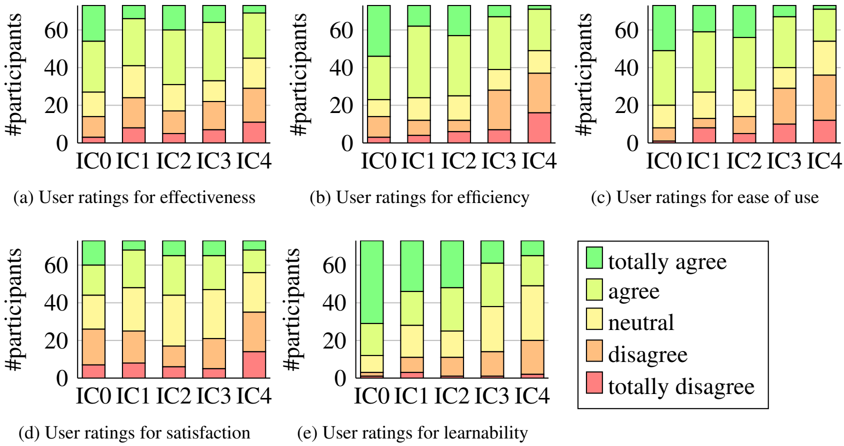
Satisfaction: “I liked using this input method.”

Learnability: “I could use this input method intuitively (without reading the description)” up to “I doubt I will ever be able to confidently use this input method (even after training).”

We additionally offered a textfield for positive and negative remarks, reasons for their judgments, further suggestions, questions, or comments. In the end, participants completed two rankings, one according to how much they liked the ICs and one on how useful they perceived the ICs to be.

Participants

We recruited prospective participants via social media and a list of volunteers maintained by the department. In total, 75 participants (34 female, 40 male, 1 preferred not to say) completed the online survey, but two male participants had



(a) User ratings for effectiveness

(b) User ratings for efficiency

(c) User ratings for ease of use

(d) User ratings for satisfaction

(e) User ratings for learnability

Figure 6.5: Results of the online survey showing the agreements of 73 participants on a five-point Likert scale about the input controls (IC0 to IC4). The exact formulation of the statements is provided in Section 6.2.2.

to be excluded from the analysis. We analyzed the data from 73 participants with an average age of 26.0 years ($SD = 6.0$). More than 90 % had a high school or higher degree. They additionally reported to on average having some knowledge about stochastics and statistics.

Results

All answers of the participants are depicted in Figure 6.5. For the analysis of the results, all Likert scale ratings were converted to the numbers 1 (totally disagree) to 5 (totally agree). The learning item was converted to the same scale. For each statement, we conducted a Friedman test with a significance level of $\alpha = 0.05$. As post hoc analysis, we conducted Wilcoxon signed-rank tests with an applied Bonferroni correction, resulting in a significance level of $p < .005$ for each statement. We only report significant results. We additionally report qualitative feedback.

Effectiveness. We found a significant difference in terms of perceived confidence to be able to correctly enter data, $\chi^2(4) = 23.94, p < .001$. Participants rated IC1 and IC4 significantly worse than IC0 and IC2.

Efficiency. We found a significant difference in terms of perceived ability to quickly enter data, $\chi^2(4) = 61.56, p < .001$. IC4 was rated significantly worse than all other input controls, and IC3 was rated significantly worse than IC0, IC1, and IC2.

Ease of Use. We found a significant difference in terms of perceived ease of use, $\chi^2(4) = 63.85, p < .001$. Participants rated IC3 and IC4 significantly worse than IC0, IC1, and IC2

Satisfaction. We found a significant difference in terms of perceived satisfaction, $\chi^2(4) = 16.48, p = .002$. IC4 was rated significantly worse than IC2 and IC3.

Learnability. We found a significant difference in terms of perceived learnability, $\chi^2(4) = 73.74, p < .001$. Participants thought that IC0 would be significantly easier to learn than all other input controls. IC3 and IC4 were also rated worse than IC1 and IC2.

Ranking - Likeability. We found a significant difference in the ranking of the input controls for likeability , $\chi^2(4) = 31.46, p < .001$. Participants ranked IC0, IC2, and IC3 significantly better than IC4. Additionally IC2 was ranked significantly better than IC1.

Ranking - Usefulness. We also found a significant difference in the ranking of the input controls for usefulness , $\chi^2(4) = 62.18, p < .001$. Participants found IC2, IC3, and IC4 significantly more useful than IC0 and IC1.

Qualitative Feedback. We collected qualitative feedback for all input controls and the ranking. The feedback about the rankings gave most insights into why participants preferred some controls over others. One participant summarized the ranking in his comments by highlighting that he "*[...] liked [IC3] because you can adjust some features, but it's still relatively quick. Method [IC2] and [IC1] give less possibilities. But in [IC4] it's too much you have to enter.*" Most participants preferred IC3 as a good compromise: "*The more power the methods hold, the more complex and annoying the interaction got. Number [IC3] was a good compromise. Easy sliding and the possibility to change the width.*"

Discussion

According to the ranking results, participants preferred IC2 and IC3 over the other methods, and IC0 over IC2 on all scales other than satisfaction. We assume that IC2 outperformed IC0 on this scale as the participants realized that IC0 was not suitable to provide an answer for the example task. Interestingly, participants also rated IC2 better than IC1 for all given statements although IC1 was easier to use and provided fewer interactive items. One reason for this preference was that participants did not like the fixed slider bar as they did not understand how the range was chosen. This resulted in IC1 having the worst satisfaction rating after IC4. To raise the satisfaction for IC1, it could help to provide a detailed explanation on why the chosen range was appropriate for a given task. Of all input methods, participants rated IC4 worst on all scales. They perceived the control as cumbersome and difficult to handle. Although participants did not rate IC3 high on the single scales, it was placed very high in the rankings and participants stated in their comments that it was the best compromise between complexity and possibilities.

6.2.3 Evaluation in the Lab

As a second evaluation, we conducted a user study in the lab. In this study, we concentrated on the input time, additional support to understand the controls, and how well participants provided the expected input when using the controls.

Method

We used a within-subjects design. Each participant had to solve three tasks and complete a SUS questionnaire for each input control. To minimize sequence and learning effects, we randomized the order of the input controls. Each participant completed the following three tasks per input control:

1. The task was an adapted version of the exemplary question of the online evaluation. The question was slightly modified to "What is the most likely value. . . ?" We also prepared five different tables and randomly assigned them to the input controls for each participant.

2. Participants had to specify how much money they spent when doing their grocery shopping. This was deliberately chosen as a free task with no right answer. Participants should get a feeling on how they would interact with the input control in a real task in everyday life.

3. Participants had to specify the possible outcome of dice rolls.

The study interface was a web-based interface that contained the description of the task, the question, the input control, a help button, and five buttons for participants to judge their confidence in the correctness of their answer with

Table 6.5: Results of the user study showing the means and standard deviations of all 30 participants for each of the used metrics.

	IC0		**IC1**		**IC2**		**IC3**		**IC4**	
	M	SD	M	SD	M	SD	M	SD	M	SD
Clicks per Element	1.86	1.63	2.89	2.81	1.43	0.98	1.84	1.72	2.53	1.56
Total Input Time (s)	33.2	32.4	35.5	30.1	31.4	21.8	45.6	33.7	63.3	32.5
Help Time (s)	8.90	26.67	5.71	17.24	5.71	17.24	7.18	27.49	10.99	27.81
Perceived Correctness	3.44	1.06	3.49	1.11	3.82	1.03	3.91	0.83	3.76	0.87
Deviation of Answers T1	2.37	1.90	2.63	1.97	2.59	2.08	2.64	2.14	2.39	1.82
Deviation of Answers T3	53.17	85.00	42.17	36.31	29.44	54.47	40.56	62.91	30.78	46.9
SUS score	65.25	20.05	65.67	16.00	70.92	17.43	72.5	16.19	47.83	21.52

five options: "*My input is correct*", "*My input is nearly correct*", "*I'm not sure whether my input is correct*", "*I doubt that my input is correct*", and "*I don't understand it*". We recorded all clicks on the input controls and buttons, the time the help dialog was open, the total input time, and the input itself.

Participants

In total, we conducted the lab study with 30 participants (11 female, 19 male) with an average age of 26.0 ($SD = 7.2$). We recruited participants in a university setting, so the majority of them were undergraduate students with different subjects. As in the online evaluation, the majority had a high school degree as their highest educational degree, thus a similar amount of knowledge about stochastics and statistics.

Results

For the analysis, we did not distinguish between participants with low or high statistical knowledge as a calculation of Pearson's r to detect possible correlations between the statistical knowledge of participants and their performance revealed no strong positive or negative correlations. We only found four moderately positive correlations ($0.40 < r < 0.45$) in terms of how often and how long participants consulted the help for tasks 2 and 3 of IC2 and IC4. All other correlations were not significant ($-0.35 < r < 0.35$). As for the online evaluation, we conducted Friedman tests with a significance level of $\alpha = 0.05$ and Wilcoxon signed-rank tests with an applied Bonferroni correction ($p < .005$) for post hoc analysis. Table 6.5 depicts means and standard deviations for all metrics.

Efficiency. We recorded all clicks on the input controls to understand how often participants changed their input. To calculate a comparable value, we divided the total number of clicks for one input control by the number of interactive elements (drag handles). The Friedman test showed a significant difference in how often participants clicked on the input controls, $\chi^2(4) = 43.54, p < .001$. Most clicks per element were recorded for IC1 with 2.89 clicks, which was significantly higher than for IC0, IC3 and IC2. IC1 was followed by IC4, which was also clicked significantly more often than IC2 and IC3. We additionally analyzed the total input time. The Friedman test showed a significant difference in how long the participants interacted with the input controls, $\chi^2(4) = 98.05, p < .001$. In comparison, participants were faster using IC0, IC1, and IC2. With IC3, it took them significantly more time to enter data than with IC0 and IC2. Entering data with IC4 took significantly longer than with all other input controls.

Confidence. We recorded how often participants clicked the help button and how much time the help information was opened. We did not find any significant difference for the usage of the help button and the time. At the end of each task, participants additionally rated their confidence of the correctness of their input. We converted the statements to numbers from 1: “*I don’t understand it*", to 5: “*My input is correct*". The Friedman test showed a significant difference in the confidence, $\chi^2(4) = 16.85, p = .002$. Participants were significantly more convinced of the correctness of their input when using IC3 than when using IC0 and IC1.

Effectiveness. We calculated the absolute deviation of the answers for task 1 and task 3 by calculating the mean deviation of all interactive elements, but the Friedman test showed no significant differences.

Usability. Each participant answered a SUS questionnaire for each input control which was adapted by replacing the term *system* with the term *input control*. We found a significant difference in the reported usability, $\chi^2(4) = 33.64, p < .001$. IC4 was rated significantly worse than all other input controls with a SUS score of 47.83.

Discussion

The results indicate that IC2 was easier to use intuitively than IC0 and IC1, as the least clicks per interactive element were recorded for IC2. Additionally, participants were faster with IC2 than with IC0 and IC1, although IC2 consisted of one more interactive control than the other two ICs.

We additionally experienced that help information was used surprisingly often for IC0, the standard slider. We assume that participants were unsure on what value to enter as the control did not allow them to specify uncertainty. The help time for IC0 was also longer than for all other ICs.

Regarding confidence, participants were mostly confident of the correctness of their answers for IC3 and IC2, although analyzing the effectiveness showed that IC3 had the highest deviation from the exact value. The highest standard deviation could be a sign that participants with lower statistical knowledge had problems to actually understand the IC. Nevertheless, the SUS score of IC3 was the highest, followed by the score for IC2.

6.2.4 Implications

Based on the results of the survey and the lab study, we derived implications and recommendations for the usage of the five input controls.

Basic Slider (IC0): The basic slider can be used if the other sliders are not applicable or if it is important to allow fast input. However, an ordinary number field might lead to better results than using a slider. If the users are uncertain about the expected input, they might be confused by the restriction of a single value interface. To avoid confusion, it is very important to state whether mean, mode, or median is the expected input value.

Fixed Range Slider (IC1): The results for the fixed range slider were neither promising in the online survey, nor in the lab study. Participants were unsatisfied with the interaction and unconfident about the fixed size of the range, which did not necessarily match their expectations. We suggest to preferably use the flexible range slider, except for cases in which the fixed range can be easily explained to users.

Flexible Range Slider (IC 2): The flexible range slider is applicable in a wide range of scenarios and even people without a huge amount of statistical knowledge felt confident in using it to correctly enter data. Specifying the minimum and maximum value seemed to be an intuitive task. The studies showed that participants were confident about their input, and usability ratings were high.

Flexible Range Best Estimate Slider (IC 3): The flexible range best estimate slider provides a good compromise between functionality and understandability. Participants stated that they preferred using it, which is also supported by the high ranking for likeability and usefulness in the online survey and the good usability score in the user study. However, participants did not necessarily enter correct data, which suggests that it is already difficult to use for people with less statistical knowledge. Basic statistical knowledge is a prerequisite for correctly using the slider.

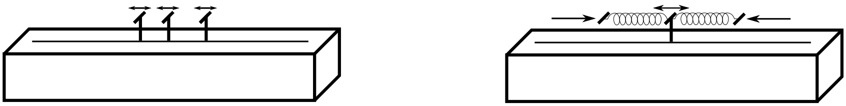
Advanced Flexible Range Best Estimate Slider (IC 4): The advanced flexible range best estimate slider was mainly judged negatively by participants. This is also reflected in the bad SUS score that the control received in the user study. We therefore do not recommend to use this input control for the general public. It may, however, be useful for people with high statistical knowledge. Participants with high statistical knowledge appreciated the degrees of freedom that the control offers.

Overall, we found that all input controls are suitable to use in different contexts and for different types of users. However, IC2 is the most promising candidate to fit most contexts well.

6.3 Tangible Shape-Changing Input Controls

Tangible and shape-changing interfaces offer promising new ways of interaction. A tangible interface, also known as graspable interface, allows for physical manipulation [Fitzmaurice, 1997]. Shape-changing interfaces additionally allow the deformation from one shape to another [Rasmussen et al., 2012]. Roudaut et al. [2013] provide a taxonomy of the different types of shape-change: area, granularity, porosity, curvature, amplitude, zero-crossing, closure, stretchability, strength, and speed that we used as basis for our prototype brainstorming. Shape-changing interfaces seem promising for the input of uncertain data as the shape-change aspect could be used for the transfer between a certain and an uncertain input.

The main goal of this part of the work is to understand whether tangible shape-changing interfaces are suitable for the input of uncertain data. In the following, we present six possible designs for shape-changing interfaces that could be used for the input of uncertain data. With the help of focus groups, we selected the most promising design and built a fully functional prototype. The prototype was then evaluated in a lab study.



(a) Split Slider

(b) Spring Slider

Figure 6.6: Concepts of the split slider and the spring slider.

6.3.1 Prototype Designs and Qualitative Evaluation

We developed six possible designs for shape-changing tangible input controls that allow the input of uncertainty. We based the designs on the taxonomy by Card et al. [1991] distinguishing between slider-based and dial-based designs and used the taxonomy by Roudaut et al. [2013] to brainstorm different ideas. In the following, we briefly describe all designs.

Slider-Based Designs

Sliders are common input controls in graphical and tangible user interfaces. As described in Section 6.2, they are suitable for the input of uncertainty due to many advantages that they offer in comparison to other input controls.

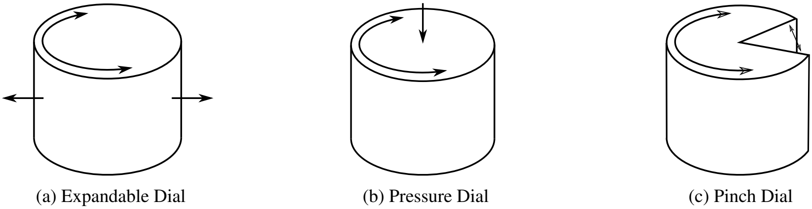
Split Slider. The split slider (see Figure 6.6a) represents a standard slider with a splittable knob. In one-knob mode, the slider allows users to enter a deterministic value. In multi-knob mode, the slider allows users to enter a range and a best estimate corresponding to a probability distribution function. The design resembles the flexible range best estimate slider from Section 6.2.

Spring Slider. The spring slider (see Figure 6.6b) also works like an ordinary slider. The knob can however additionally be pinched together. A completely pinched together knob resembles a very certain input, while applying less strength resembles an uncertain input.

Speed Slider. The speed slider incorporates the speed in which the user moves the knob of the slider as an input. The faster the user selects an input, the more certain the user is.

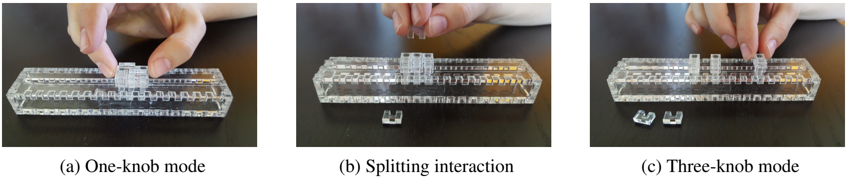
Dial-Based Designs

Similar to sliders, dials are commonly known input controls, especially in the tangible form used for voice control. Dials are promising for uncertainty input as



(a) Expandable Dial (b) Pressure Dial (c) Pinch Dial

Figure 6.7: Concepts of all three dial-based designs.



(a) One-knob mode (b) Splitting interaction (c) Three-knob mode

Figure 6.8: First non-functional prototype for the split slider.

they have several degrees of freedom that can be used in addition to the rotation to allow a user to enter uncertainty. For all dial-based designs, the input value can be selected by rotation whilst additional parameters allow a user to input uncertainty.

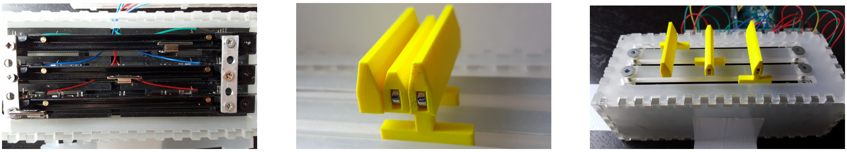
Expandable Dial. The expandable dial (see Figure 6.7a) can be increased or reduced in size in addition to its rotary interaction. A more extended dial resembles a more uncertain input.

Pressure Dial. The pressure dial (see Figure 6.7b) can be pressed downwards to enter an additional value for uncertainty. The stronger/further the dial is pressed down, the more certain the input is.

Pinch Dial. The pinch dial (see Figure 6.7c) has an open space than can be pinched together. The more the dial is pinched, the more certain the input is.

Qualitative Evaluation

We used existing low-fidelity prototypes for the dial-based designs and additionally built low-fidelity prototypes for the split slider (see Figure 6.8) and the spring slider. The low-fidelity prototypes were built out of paper or plastic and cut with the help of a laser cutter. All low-fidelity prototypes were non-functional.



(a) Inside view of the slider (b) Close up of the one-knob mode (c) Three-knob mode

Figure 6.9: Functional prototype of the split slider.

We conducted two focus groups with in total 12 participants (10 male, 2 female), who had an average age of 24.9 ($SD = 3.7$). Participants first developed scenarios in which they would like to use uncertain input and ranked them according to their preferences. In pairs, they then picked one scenario and got a prototype which they had to evaluate in terms of suitability for their scenario. Participants preferred the split slider and stated that it was intuitive and flexible. They appreciated the simple design and the visual feedback. In general, they disliked the dial-based designs for the uncertainty input being disconnected from the actual input. In the following, we therefore selected the split slider as the most promising candidate design to build a fully functional prototype and conduct a lab study.

6.3.2 Evaluation in the Lab

We conducted an explorative user study in the lab to understand whether our slider design would be suitable for the input of uncertain data. For the study, we implemented a functional prototype of the split slider (see Figure 6.9) by using three slider potentiometers and an Arduino. We created the illusion that the prototype was one slider with three knobs by enlarging the single knobs. The knob was designed to allow for one-finger control as well as easy splitting. Magnets gave users a haptic feedback when splitting the knob.

Method

Each participants had to answer twelve questions with the help of the slider. We constructed the questions to match a public survey in a train with continuous answer scales. For example, for the question “How often do you use the train?”, participants answered the question on a scale from “Never” to “Daily”. We used a 12x12 Latin square design to randomize questions across participants. The study consisted of two phases. In phase one, participants had to answer the first half of the questions without any explanation of the slider. Before answering the second half of the questions in phase two, they received detailed instructions on how the slider works. After both phases, participants had to fill in a questionnaire which consisted of the UMUX questionnaire [Finstad, 2010], one question about the suitability of the prototype for uncertain input, and one question on the understanding of the prototype. All questions consisted of a seven-point Likert item. We additionally measured the input time and stored knob positions and movements.

Participants

In total, 18 participants (13 male, 5 female) with an average age of 34.4 ($SD = 14.9$) participated in the study. Six of them had previous knowledge of the prototype, whilst the other twelve participants had never seen or heard about the prototype before. In the following, we only analyze the data from the participants that had never seen the prototype before.

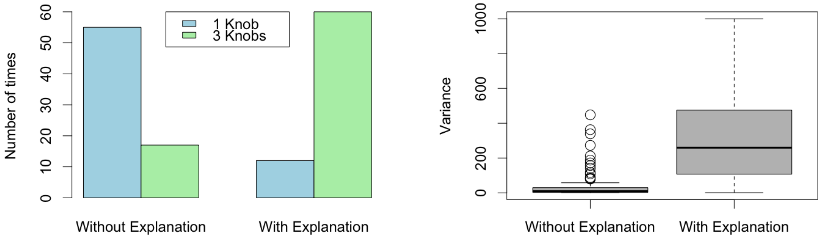
Results

We analyzed the UMUX, the additional questions, as well as the usage behavior of the participants. We therefore converted all Likert items to the numbers 1 to 7, where 1 corresponds to "strongly disagree" and therefore a negative answer, and 7 corresponds to "strongly agree" and a positive answer.

Usability. Overall, our prototype received an UMUX score of 82.4 ($SD = 17.00$), which for the SUS according to [Bangor et al., 2008] corresponds to an excellent score. A Friedman test did not show any significant differences for the effect of knowledge (phase 1 vs. phase 2) on the usability, $\chi^2 = 0.818, p = 0.366$.

Prototype Suitability and Understanding. Without explanation, the suitability of the prototype for uncertain input was rated slightly positive ($M = 4.0$, $SD = 2.7$). A Friedman test showed that after the explanation, this value increased significantly ($\chi^2(1) = 7.0, p < .01$) to an average of 6.8 ($SD = 0.4$). The perceived understanding of the prototype also increased slightly ($\chi^2(1) = 5.0, p < .05$) from 6.2 ($SD = 1.5$) to 6.9 ($SD = 0.3$).

Splitting Behavior. Figure 6.10a shows how often participants used the different knob modes. A chi-square test of independence showed that participants used the three-knob mode significantly more often in the second phase after receiving the explanation of the prototype, $\chi^2(1) = 49.237, p < .001$. Similarly, the expressed variance (distance from the outer knobs, see Figure 6.10b), increased drastically



(a) Knob mode usage per phase

(b) Expressed variance per phase

Figure 6.10: Influence of the explanation of the prototype on the usage of different knob modes and the expressed variance for participants without previous knowledge of the prototype.

from phase 1 ($M = 47.4$, $SD = 90.3$) to phase 2 ($M = 303.4$, $SD = 249.2$). An independent 2-group t-test revealed that this difference is statistically significant, $t(89.352) = -8.1942, p < .001$.

Input Time. Participants needed on average only half of the time to give a deterministic answer ($M = 12.0\,s$, $SD = 7.3\,s$) in contrast to giving an answer with uncertainty ($M = 21.2\,s$, $SD = 11.6\,s$). We found a weak correlation between the expressed variance and the input time, $r(214) = 0.345, p < .001$. However, the input time might drop with further usage. We also found a weak correlation between task order and input time (learning effect) for phase 1 ($r(106) = -0.293, p < .01$) and phase 2 ($r(106) = -0.295, p < .01$).

6.3.3 Discussion & Implications

Participants used the prototype very differently before and after the explanation, which lead us to the conclusion that the prototype was not self-explanatory. This problem might be only temporary, because if uncertain input becomes more common in the future, users will know how to use such devices, but for now an extra explanation is necessary for people to understand the possibilities of the input device. However, the ratings of the UMUX score show that our prototype is very well usable in both modes, one-knob and three-knob mode. This indicates that people can use the split slider in the standard way if they do not know about the extra functionality without being confused. New input devices allowing uncertain input should follow this rule.

Additionally, most participants found the split slider suitable and understandable before they got the extra explanations. These values increased even more after the explanation. This also shows that participants were able to use the prototype without being confused by new possibilities. Participants tended to stick with their previous knowledge about deterministic input and mostly did not experiment with the prototype although they were invited to do so. After the explanation, participants made high use of the new functionality appreciating the split slider as an input control to enter uncertainty. This indicates that tangible shape-changing interfaces and in specific the split slider are suitable devices for uncertain input and can help users to express their uncertainty.

The results on the input time show that users needed more time to enter values with higher variance as they potentially had to move more knobs, but also think longer about the answer. This finding supports the design of allowing the standard input with a one-knob mode and the more complex input with the three-knob mode. If time constraints apply, the standard method can be used instead of the complex method. Nevertheless, further training and familiarization might reduce the time users need to enter uncertainty. However, this would need to be subject of a future study.

6.4 Physiological Sensing

Instead of offering explicit uncertainty input to a user, a system could automatically detect how uncertain a user is when providing an input. Physiological sensing can be used as input modality to substitute or accompany manual uncertainty input. Heart rate tracking is already used to understand the training performance of athletes [Tholander and Nylander, 2015], but it can also be used in interactive systems to detect the level of cognitive stress of a user. Based on the heart rate variability, interfaces can adapt their complexity accordingly to reduce stress [McDuff et al., 2016]. Besides heart rate, gaze patterns and other related gaze properties already serve as input for interactive systems when rating pictures in photo albums [Walber et al., 2014], enhancing tutoring systems [D'Mello et al., 2012], or assisting with translations [Hyrskykari et al., 2003]. Studies on gaze characteristics with the help of quiz questions [Copeland and Gedeon, 2013], problem solving tasks [Madsen et al., 2012], and language reading tasks [Karolus et al., 2017] revealed that these characteristics change when users are confronted with unfamiliar content or difficult tasks.

The main goal of this strand of research is to understand whether physiological sensing can help to detect users' uncertainty when entering data. In the following, we first describe how we generated a set of quiz questions for our user study. We then present the method and results of our user study with 24 participants who had to answer the selected quiz questions. During the study, we collected heart rate, eye tracking, and key logging data.

6.4.1 Question Selection Process

To identify suitable questions for the user study, we first built a pool of questions in a three-step selection process. In total, we transcribed 1770 German quiz questions from four books containing questions about general knowledge [Bauer and Kneip, 2016; Hotz, 2013, 2014; Pfersdorff and Glahn, 2015]. As these questions were multiple-choice, we deleted all questions that were not solvable without the multiple-choice answers and removed the answers. On the remaining 1164 questions, we applied three filter criteria to improve comparability:

- **Maximum of four words per answer:** We eliminated all questions with answers consisting of more than four words to avoid full-sentence answers. This minimizes the confounding uncertainty resulting from the need to spell long or complex phrases.

- **Maximum of 15 words per question:** We eliminated all questions consisting of more than 15 words as this is the upper border for the recommended sentence length in German for a sentence to be easily comprehensible [Seibicke, 1969]. Thus, we removed questions that could be difficult to read and comprehend.

- **Flesch-Reading-Ease of questions between 60 and 80:** The Flesch-Reading-Ease (FRE) [Flesch, 1948] is a readability metric that, based on the average sentence-length (ASL) and average syllables per word (ASW), measures how difficult is it to understand a text. We use the version for German language proposed by Amstad [1978] ($FRE_{\text{german}} = 180 - ASL - (58.5 * ASW)$) to eliminate questions that are either very easy or very hard to read in comparison to the other questions in the question set. This also minimizes the confounding uncertainty introduced by questions with different degrees of difficulty to read. We chose the interval of the FRE value to be between 60 and 80, which correlates to the category of medium easy and medium text.

Applying all these filter criteria, we reduced the question set to a number of 251 questions. To categorize these questions further, we conducted an online survey. Each page of the survey contained one question and a textfield to answer the question. Additionally, participants had to indicate how certain they were about their answer on a five-point Likert scale from “totally disagree” to “totally agree.” We randomized the order of questions per participant. In total, 59 participants provided 7,939 answers ($M = 134.6$ questions per participant, $SD = 102.7$). Based on the results of the survey, we assigned each question to one of five uncertainty classes corresponding to the item on the Likert scale that the majority of participants selected for this question. We then calculated the ratio of how often the question was rated in its respective class to the total number of answers for the question prioritizing questions with higher ratios. From the easiest and the most difficult class, we picked the 40 questions with the highest ratio; from all other classes we picked the 20 questions with the highest ratio. This resulted in a question set of 140 questions in total.

6.4.2 Evaluation in the Lab

We used the question set in a lab study to examine the relationships between users’ uncertainty and their physiological signals.

Method

Participants had to answer 140 questions of different difficulty levels (see Subsection 6.4.1), then report their perceived uncertainty for each given answer.

After arriving, participants filled in a consent from and a demographic questionnaire. They were then seated in front of a 22 in. LCD display, which showed a website with the questions running in a browser. Before starting, we attached the 3 ECG-electrodes from a NEXUS 4 to record ECG signals and calibrated the stationary eye-tracker (SMI RED 250), which was attached to the bottom of the display. The whole setup was shielded by three white plain wall constructions to avoid the disturbance of the participant by the study instructor or the appearance of the study room. After the attachments of the electrodes and the calibration phase, we gave participants a detailed explanation of the task and asked them to always provide a reasonable answer (e.g. stating a city name if the question asked for a city) and that they could take as much time as they wanted to answer a question. After a training question to get used to the study interface, they answered all 140 questions. The questions were presented one by one in a randomized order. After answering all questions, participants rated their uncertainty for each question based on a five-point Likert scale from “Strongly agree” to “Strongly disagree” to the statement: “I am sure that my answer is correct.” We again randomized the order of the questions. Participants, however, saw the question and their given answer to make the assessment. We did not ask participant for their uncertainty in between questions as this would have probably revealed too much about the context of the study.

Throughout the study, we collected data from the eye-tracker, the NEXUS, and the browser. Eye movements were recorded at $250Hz$, and the ECG signal at $256Hz$. From the browser, we collected all key events, click events, mouse movements, field focus events, completion times, and the participants’ answers. From the collected data, we derived several metrics. From the ECG signal, we calculated heart rate and heart rate variability. From the browser data, we extracted features concerning time (such as completion time, time before typing starts, time between first key stroke and last key stroke), typing behavior (such as typing speed, key down time, number of deletions), and mouse evens (such as number of clicks; length of the mouse path). From the recorded eye movements, we extracted features related to the count, time, and velocity of fixations, saccades, and blinks.

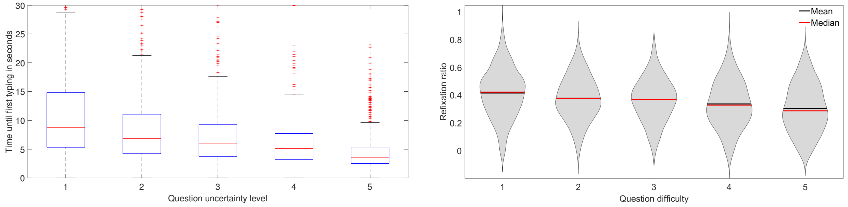
Participants

We recruited 24 participants (15 male, 9 female) with an average age of 23.2 years ($SD = 3.4$). All of them were native German speakers. Due to technical difficulties, we had to exclude data from four participants whose physiological measurements could not be tracked reliably. We also only used a subset of data for the analysis of eye movements and the heart rate due to unreliable tracking caused by make-up and loosened electrodes.

Results

We analyzed all collected data, but only present statistical results on a subset of the measures. We focus on the aspects that we identified as highly promising for detecting uncertain user input or as interesting lessons learned.

Key Logging Data. First, we investigated the time that elapsed before participants started typing an answer (see Figure 6.11a). On average, participants started typing after 9.19 s ($SD = 9.45\,s$). For the lowest self-perceived uncertainty, participants spent the lowest time until starting to type ($M = 5.15\,s$, $SD = 4.56\,s$). In contrast, they took the longest time before starting to type when they perceived



(a) Boxplot for time elapsed until participants started to type grouped by the perceived uncertainty of the answers (1: very uncertain, 5: very certain).

(b) Violin plot of the refixation ratio grouped by the perceived uncertainty of the answers (1: very uncertain, 5: very certain).

Figure 6.11: Graphical results for one key logging and one eye tracking metric.

their answer as uncertain ($M = 13.14\,s$, $SD = 13.07\,s$). A one-way ANOVA revealed a statistically significant difference ($F_{4,2717} = 84.01$, $p < .001$) in the time before typing depending on the perceived uncertainty of the answer.

Second, we investigated the time that elapsed between the first and the last key stroke of participants. On average, participants needed $4.43\,s$ ($SD = 7.13\,s$) from their first key stroke to their last key stroke when entering an answer. Participants were faster ($M = 3.73\,s$, $SD = 5.60\,s$) when their perceived uncertainty was low, while longest typing periods ($M = 5.51\,s$, $SD = 9.91\,s$) occurred when they were uncertain. A one-way ANOVA revealed a statistically significant difference ($F_{4,2717} = 8.34$, $p < .001$) in the time between first and last key stroke depending on the perceived uncertainty of the answer.

Eye Tracking Data. First, we investigate how long participants spent looking at the answer field normalized by the total time looking at the screen. On average, the ratio was 0.30 ($SD = 0.19$). When the perceived uncertainty was high, the ratio was higher ($M = 0.34$, $SD = 0.21$) than when the perceived uncertainty was low ($M = 0.26$, $SD = 0.18$). A one-way ANOVA revealed a statistically significant difference ($F_{4,1907} = 13.53$, $p < .001$) in the ratio of looking at the answer field depending on the perceived uncertainty of the answer.

Second, we investigated the amount of refixations normalized by the number of fixations for the respective question (see Figure 6.11b). On average, the ratio was 0.36 ($SD = 0.17$). When the perceived uncertainty was low, the refixation ratio was lower ($M = 0.30$, $SD = 0.17$) than when the perceived uncertainty was high ($M = 0.41$, $SD = 0.18$). A one-way ANOVA revealed a statistically significant difference ($F_{4,1907} = 23.03$, $p < .001$) in the ratio of looking at the answer field depending on the perceived uncertainty of the answer.

ECG Data. We investigated median heart rates (HR) and heart rate variability (HRV) during the study. We applied multiple aggregation strategies as a reaction in the heart rate will most likely be delayed to determine whether users' uncertainty about their answers influenced the heart rate. We used combinations of four different lag values: 0 ms, 1000 ms, 5000 ms, 9000 ms and three different window sizes: 1000 ms, 5000 ms, 10000 ms resulting in twelve different combinations. One-way ANOVAS did not reveal any significant effects of the users' uncertainty on median HR and HRV ($p > .05$).

6.4.3 Discussion & Implications

Different physiological sensors can detect uncertainty of users. In our study, a combination of key logging and eye tracking proved to be the most reliable metric to understand how uncertain users are when providing their answers. However, this could differ from context to context. Other contexts not related to quiz questions should be evaluated to better understand whether the results are comparable.

The results of our study also indicate that uncertainty does not have a significant effect on heart rate signals. This was surprising as other work indicated that heart rate is correlated with cognitive stress. We believe that the results can be explained by two reasons. First of all, wrong answers did not have any implications for participants during our study. They probably felt safe and therefore uncertain answers did not produce physiological reaction. Second, heart rate reactions might have been too slow and taken longer than the time needed for answering one of the provided questions.

We also experienced that current technology for measuring physiological signals is still prone to technical issues. Data from multiple participants was either lost or could not be tracked due to individual differences such as glasses, make-up, skin thickness, etc. Future sensors might be much better in terms of their sensing capabilities and not prone to these errors. This might lead to better and richer results.

Table 6.6: Developed input controls for uncertain data categorized by method, input modality, and uncertainty value.

Method	Input Modality	Uncertainty Value	Concrete Suggestion
Explicit	*User Interface*	Probability Percentage	Additional number field or slider
		Range	Flexible Range Slider
	Tangibles	Range	Split Slider
Implicit	*Physiologial Sensing*	Probability Percentage	Eye Tracking & Key Logging

6.5 Insights for Quantifying Uncertainty in User Input

In this chapter, we developed and evaluated multiple input controls and methods for quantifying uncertainty in user input. Table 6.6 provides an overview about the best approaches. The input methods can be categorized by whether it is an explicit or implicit method to quantify uncertainty in user input.

Explicit methods allow the user to enter a value for the uncertainty included in an input. In our work, we explored paper prototypes for qualitative methods, however they have two main disadvantages in comparison to quantitative methods. First, qualitative methods are very difficult to interpret. The interpretation can vary by user or even by system if natural language needs to be processed. Second, participants in our interviews raised the concern that qualitative input methods are very cumbersome to use. For these two reasons, we focused on quantitative methods. For digital user interfaces, we recommend using a number field or slider for additional input of a probability percentage (see Section 6.1) or the Flexible Range Slider for range input (see Section 6.2). As a tangible input control, we recommend the Split Slider (see 6.3). We additionally recommend that future input controls focus on offering two ways of interaction; the standard interaction and an additional interaction to specify uncertainty. Both interactions should be aligned to each other, but the additional interaction to enter uncertainty should not confuse the user or complicate the standard input. Thus, additional knowledge or time would not necessarily required to make a more complex input including uncertainty.

Implicit methods are methods that automatically track and quantify users' uncertainty. In Section 6.4, we explored using physiological sensing and behavioral measurements to quantify uncertainty. We suggest a combination of eye tracking data and key logging data to implicitly detect uncertainty.

In the next chapter, we focus on another important aspect of understanding how to deal with uncertainty in interactive systems: output methods. We analyze the current communication of uncertainty in mobile application and look into uncertainty visualization for step tracking and decision-making under uncertainty.

Chapter 7

Output Methods

We explored the design space of output methods for uncertain data by conducting one analysis of existing mobile applications and two individual research probes. For the analysis of existing mobile applications, we looked at what methods applications presenting uncertain data use and whether they communicate the uncertainty. We additionally conducted an online survey to understand what users in general think about such applications. For the two individual research probes, we first developed an Android application that visualized uncertainty for activity tracking data, and second developed a Facebook game to explore how players make decisions based on different amounts of uncertainty information presented to them.

The main goal of this chapter is to understand the current state of output methods for uncertain data and how uncertainty visualizations can be used to support decision-making. We used different sets of existing output methods in our studies to make a valuable comparison.

Parts of this chapter are based on the following publications:

- M. Greis, T. Ohler, N. Henze, and A. Schmidt. Investigating representation alternatives for communicating uncertainty to non-experts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9299, pages 256–263, 2015.
- M. Greis, P. E.Agroudy, H. Schuff, T. Machulla, and A. Schmidt. 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*, pages 2–5, 2016.

7.1 Communication of Uncertainty in Current Mobile Applications

Many mobile applications show uncertain data such as weather forecasts, location data, or measurements of physical activity. However, it is not clear whether or how they present uncertainty to their users and what methods they use to generate trust in their measurements and predictions. Additionally, users might prefer a different communication that has not so far been used in mobile applications.

The main goal of our work is to understand the methods current mobile applications use to communicate uncertain data and whether users are satisfied with this communication. We analyzed 30 mobile applications in the three areas of weather forecasting, navigation/fuel prizes, and healthcare. We picked the most installed applications available for free in the Google Playstore and the Apple Appstore. We were mainly interested in which methods these applications use to show data and whether they indicate the uncertainty of the data. We additionally conducted an online survey to assess users' trust in the reliability of such applications and their preferences for the communication of uncertainty.

In the following, we present the analysis of 10 weather applications, 10 navigation/fuel prize applications, and 10 healthcare applications. We mainly focus on

Table 7.1: Methods used to display uncertain data in the following weather applications: W1 - AccuWeather, W2 - BayWa Agri-Check, W3 - Blitzortung Gewitter-Monitor, W4 - Regenradar, W5 - wetter.com, W6 - wetter.info, W7 - Wetter 14 Tage, W8 - Wetter Deutschland XL PRO, W9 - Wetter Online, and W10 - Windfinder.

Methods	**W1**	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**
Diagrams	✓	✓			✓	✓				
Text	✓	✓					✓	✓	✓	
Symbols	✓	✓	✓			✓	✓	✓	✓	✓
Percentages	✓	✓			✓	✓	✓	✓	✓	✓
Values	✓	✓	✓		✓	✓	✓	✓	✓	✓
Maps	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

how the information in these applications is depicted. We additionally present the results of our online survey.

7.1.1 Analysis of Mobile Applications

We downloaded ten free weather applications, ten free navigation/fuel prize applications, and ten free healthcare applications from the Google Playstore and Apple Appstore. We focused on applications available in German. We then took screenshots of all different screens of the applications to categorize how the information was displayed.

Weather Applications

Table 7.1 shows an overview of the analyzed weather applications and how they displayed weather information. The applications used six different methods: Four used diagrams (e.g. line graphs, pie charts), five used verbal expressions (e.g. “low” or “high”), eight used symbols (e.g. grey clouds or a yellow sun), eight used percentage values (e.g. the probability of rain), nine used concrete values (e.g. “16 ° C”), and all used colored maps (e.g. for showing temperatures, rain intensity). Besides showing the probability of rain, none of the applications indicated the uncertainty of the information. All weather applications showed the data as deterministic values.

Table 7.2: Methods used to display uncertain data in the following navigation/fuel prize applications: N1 - Blitzer.de, N2 - Blitzer!, N3 - Clever Tanken, N4 - DB Navigator, N5 - Google Maps, N6 - iOS Karten, N7 - MyTaxi - Die Taxi App, N8 - Offline Maps Navigation, N9 - StauMobil, and N10 - Tanken App.

Methods	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**
Text	✓		✓	✓	✓	✓	✓	✓	✓	✓
Symbols		✓		✓		✓	✓	✓	✓	✓
Values		✓	✓	✓	✓	✓	✓	✓	✓	✓
Maps	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Navigation/Fuel Prize Applications

Table 7.2 shows an overview of the analyzed applications in the field of navigation and fuel prize applications. The applications used four different methods for displaying information. Seven applications used symbols (e.g. a pin for the location), nine used text (e.g. a textual description of a traffic jam prediction), nine used values (e.g. “Distance: 9.4 km”), and all used maps (e.g. to display the location and routes). One application (MyTaxi - Die Taxi App) used verbal expressions to convey uncertainty, by stating that it takes “approximately” an amount of time until the cab arrives and that the prize is an “approximate value”. Two more applications used the terms “prediction” or “forecast” in their headline, but gave no detailed information about how uncertain their information was or how it was calculated.

Healthcare Applications

Table 7.3 shows an overview of the analyzed applications in the field of healthcare. The applications used five different methods for displaying information. Two applications used maps (e.g. to show walking routes), six used diagrams (e.g. bar charts, pie charts), six used symbols (e.g. a green human, a red heart), seven used text (e.g. “in good shape”), and all used values (e.g. “103 bpm”). One application used the word “approximately” to indicate that the values were uncertain. Additionally, the application “Clue - Menstruationskalender” showed “$(+/-x$ days)” next to its predictions, indicating how much the prediction of the day was uncertain. None of the other applications showed how uncertain the presented information was.

Table 7.3: Methods used to display uncertain data in the following healthcare applications: H1 - Schrittzaehlen - Accupedo, H2 - Cardiio, H3 - Clue - Menstruationskalender, H4 - Idealgewicht, H5 - Kardio - Herzfrequenz Monitor, H6 - Komoot - Fahrrad und Wander GPS, H7 - Nichtraucher Coach, H8 - Promillerechner Live, H9 - Runtastic GPS, and H10 - Sehtest.

Methods	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
Text		✓	✓	✓		✓	✓	✓		✓
Symbols		✓	✓		✓	✓		✓	✓	
Values	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Diagrams	✓	✓	✓		✓	✓			✓	
Maps						✓			✓	

7.1.2 Online Survey

We conducted an online survey to assess users' general perception of applications in the analyzed areas. We were mainly interested in their opinions about the reliability of measurements and predictions in these applications. Additionally, we wanted to know whether they would like to get more information about the underlying uncertainty of the displayed data.

Method

We first asked participants of the online survey to provide demographic information such as gender and age. We then asked whether they use weather, navigation, or healthcare applications. Depending on what type of applications participants used, they answered the following questions for each type of application they used: (1) How reliable do you think the measurements are? (2) How reliable do you think the predictions are? (3) Would you like to see more clearly whether the data is uncertain? Participants answered all questions on a five-point Likert scale from "not at all" to "very much".

Participants

115 participants (69 male, 45 female, 1 preferred not to say) between the age of 13 and 51 took part in our online survey. Around 80 % used weather and navigation applications, and more than 50 % used healthcare applications.

Results

We converted participants' answers from the Likert items to a scale from 1 ("not at all") to 5 ("very much"). In general, we found that healthcare measurements seemed to be most unreliable for participants ($M = 3.0$, $SD = 1.0$), followed by weather measurements ($M = 3.9$, $SD = 1.0$) and measurements in navigation applications ($M = 4.2$, $SD = 0.9$). Predictions seemed to be considered slightly more unreliable compared to measurements for weather applications ($M = 3.7$, $SD = 0.9$) and navigation applications ($M = 4.1$, $SD = 0.9$). Although participants in general did not see the reliability on the negative end, participants wished to have information about the uncertainty in all applications (weather: $M = 4.2$, $SD = 0.9$, healthcare: $M = 4.0$, $SD = 1.1$, navigation: $M = 3.6$, $SD = 1.2$).

7.1.3 Discussion & Implications

The analysis of existing mobile applications showed that applications seldom communicate uncertainty information. The rare occurrence of applications communicating this information mostly use verbal expressions such as "approximately". This leaves quite some open room for future improvement and implementation of uncertainty visualizations.

The results of our survey indicate that users in general seem to find predictions more unreliable than measurements. How reliable they perceive an application to be depends on the area of the application and potentially the application design itself.

We also found that users actually wish for uncertainty to be displayed in mobile applications. Displaying the uncertainty of data could potentially help to even increase the perceived reliability of an application.

7.2 Uncertainty Visualization for Activity Tracking

Multiple studies have already showed that the measuring accuracy of activity trackers depends on the brand and kind of device [Case et al., 2015; Guo et al., 2013], the body part it is worn on [Sasaki et al., 2015], and the walking speed [Crouter et al., 2003]. The measurement errors for step counts range from 1 % up to nearly 30 % [Guo et al., 2013]. We therefore wanted to understand how this uncertainty in the measurements could be communicated to the user.

The main goal of the work presented in this section is to first understand how users perceive the uncertainty of the measurements and how much they trust their devices and second, develop graphical overviews for activity data that take into account uncertainty.

7.2.1 Online Survey

We conducted an online survey to assess how people use activity trackers and how much they know about the uncertainty of their measurements. We additionally used the online survey to determine people's beliefs on how uncertain activity trackers are and to collect ideas on how the uncertainty of activity trackers could be visualized as tracking results.

Method

We first asked participants to provide demographic data such as gender and age. We then asked some general questions about sports and activity tracking. Participants were then separated into four different groups: those that currently used activity trackers, those that had used an activity tracker in the past, those that used other devices for activity tracking (such as a treadmill), and participants that had never used any activity tracking device. We were mainly interested in participants who had already used an activity tracker or currently used one. Depending on their experience with activity trackers, participants answered different questions over the course of the study. We asked those who had used activity trackers before to answer questions about the measuring reliability, graphical overviews used by their activity trackers, and errors that they had spotted. We structured all questions as five-point Likert items. After these basic questions, participants had to answer concrete questions about how much the displayed results of an activity tracker (e.g., 2000 steps, 5000 steps, and 9000 steps) deviate from the true value. At the end of the survey, we asked participants for their opinion about visualizing uncertainty for activity tracking data and their ideas on how this could be accomplished.

Participants

In total, 364 participants (153 female, 206 male, 5 preferred not to say) with an average age of 28.4 ($SD = 8.8$) completed the online survey. Participants did all

Table 7.4: Participants' agreement on a five-point Likert scale with the presented statements.

Statement	M	SD
I am satisfied with the functionality of my activity tracker.	4.3	0.6
I trust the exactness of the measurements of my activity tracker.	4.0	0.6
I am satisfied with the graphical overview in the application.	4.1	0.8
I know that the measurements of activity trackers are not always correct.	4.4	0.9
I always remember the exact value displayed on my activity tracker.	2.5	1.1
In my thoughts, I round the values displayed on my activity tracker up/down.	2.6	1.2

kind of sports, e.g. swimming, soccer, bicycling, running, fitness training or more specialized sports such as archery or diving. 118 of our participants stated that they had never used an activity tracker before; 46 only used other devices that track activity data (e.g. a treadmill), and 29 indicated that they had stopped using their activity tracker. 171 participants (97 female, 72 male, 1 preferred not to say) were current users of activity tackers at the time of the survey.

Results

In our analysis, we only used the data provided by the 171 current users, as we did not have enough participants who had abandoned their activity tracker to get an indication about how that might affect the perceived reliability. We converted all Likert items to a number from "totally disagree" as 1 to "totally agree" as 5. The answers for the basic questions on reliability are displayed in Table 7.4. Although most of the participants stated that they knew the measurements of activity trackers were not always correct, 17.5 % stated that they had not been aware of this before taking part in the survey.

For the reliability of measurements of activity trackers, we found that for most measurements, participants thought that the measurement error would be between 2.2 %-5.6 % for steps (see Figure 7.1) and between 1.4 %-10.5 % for distances (see Figure 7.4). For the treadmill, participants estimated a higher percentage of measurement errors for the calories between 5 %-16.7 % (see Figure 7.3), however, they estimated a lower error for distances between 1.7 %-4 % (see Figure 7.4). The estimated error is in the same range for different amounts/distances. However, for the distance on the treadmill, the answers are not normally distributed as for the other three estimations.

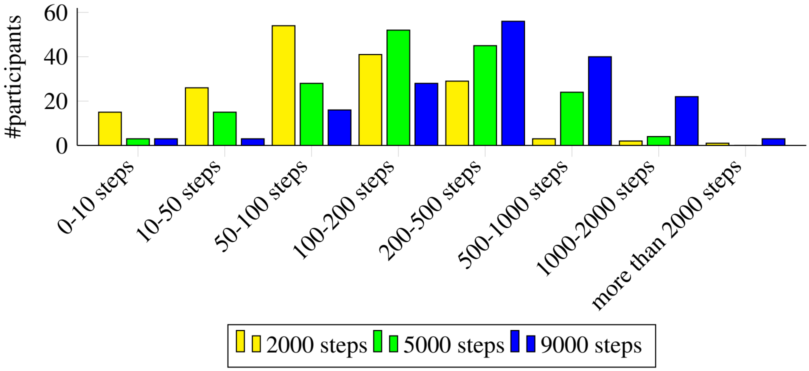


Figure 7.1: Participants' answers on the question "If an activity tracker measures ... steps on one day, how high do you estimate the error of the measurement?" for the amount of 2000, 5000, and 9000 steps.

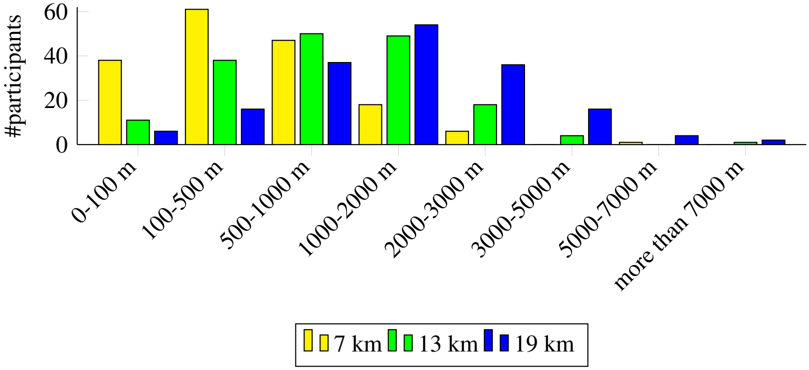


Figure 7.2: Participants' answers on the question "If an activity tracker measures a covered distance of ... kilometers on one day, how high do you estimate the error of the measurement?" for the amount of 7, 13, and 19 km.

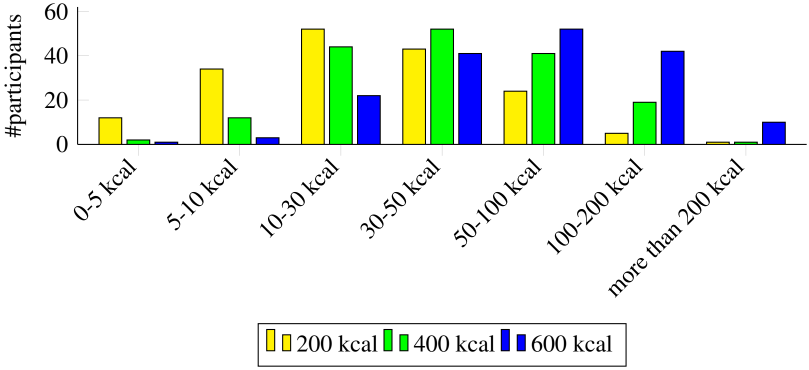


Figure 7.3: Participants' answers on the question "If a treadmill measures a calorie expenditure of ... in 30 minutes, how high do you estimate the error of the measurement?" for the amount of 200, 400, and 600 kcal.

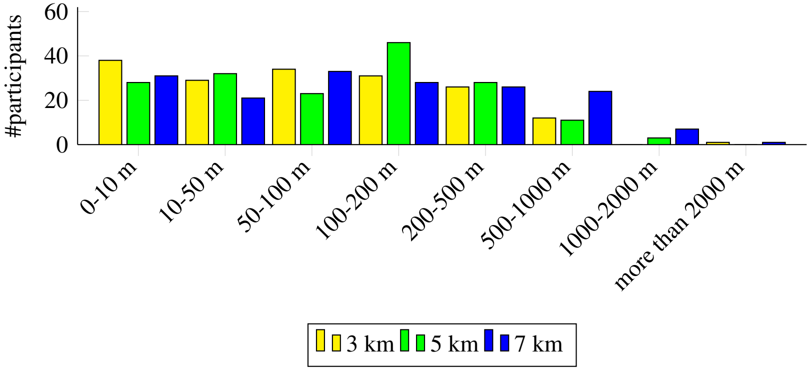


Figure 7.4: Participants' answers on the question "If a treadmill measures a covered distance of ... kilometers in 30 minutes, how high do you estimate the error of the measurement?" for the amount of 3, 5, and 7 km.

Table 7.5: Participants' agreement on a five-point Likert scale with the presented statements.

Statement	M	SD
If my activity tracker shows 492 kcal after my training, I would continue my training to reach 500 kcal.	3.4	1.4
The error of measurement of activity trackers is that high that the results of step and calorie tracking are too unreliable.	2.7	0.9
I would prefer if manufacturers of activity trackers would publish more information about possible errors of measurement.	3.8	0.9
I would like an activity tracker to take into account potential errors of measurements in the graphical overview.	3.6	1.1

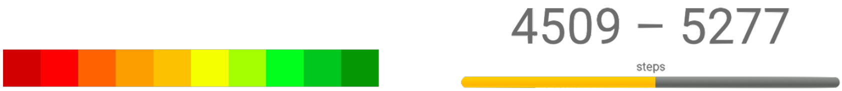
Although participants stated that they knew the measurements of activity trackers might not be completely reliable, they still slightly agreed with the statement that they would continue their training if their device showed 492 kcal (see Table 7.5). They did not find their devices too unreliable, but would prefer to have more information on the possible errors of measurements and also stated a preference for having them displayed in the graphical overview.

Participants had several ideas on how the measurement uncertainty could be displayed in the graphical overview of their activity trackers. They for example suggested to show an error percentage, a minimum and maximum value (e.g. grey bar, boxplot), a color as indication how uncertain the value is, or an overview that only roughly marks the value instead of providing a single value.

Discussion

Although most users know that the data displayed by activity trackers is not completely reliable, there are users who seem to trust the exact value shown on the display. Additionally, although most users know that the activity data is unreliable or could be wrong, they would continue training to reach a certain threshold (e.g. 500 kcal). We assume that psychological aspects play an important role in making users want to reach the threshold despite the knowledge that their actual calorie expenditure might differ from the displayed amount.

Participants estimated the error of measurement for steps up to 5.6 %. The real error of measurement, however, can be much higher depending on the used device. A group of participants also seemed to trust a treadmill more in terms of distance than an activity tracker. This can be seen in the unusual distribution of the data. Participants also applied different estimates for different data types, for example



(a) Color palette

(b) Range of values

Figure 7.5: To develop graphical overviews for activity tracking data, we created a standard color palette to indicate whether the users reached their goals and additionally replaced the normal deterministic value with a displayed range.

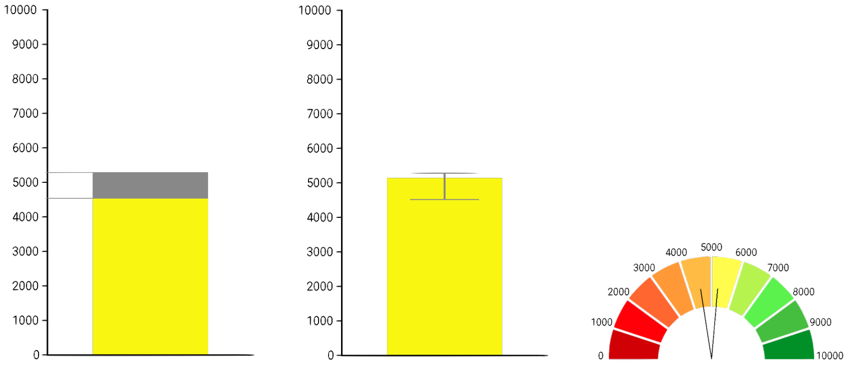
steps and calories. From these findings, we can learn that it depends on the actual type or maybe even the brand of a device and the data type who uncertain participants estimate measurements and predictions.

We also found that participants in general would like to have more information about errors of measurement. They also expressed a slight agreement for displaying uncertainty in the graphical overview and had diverse suggestions on how the graphical overview could be adapted to fulfil this requirement.

7.2.2 Evaluation in the Wild

Based on participants' suggestions in the online survey, we developed two basic components for the display of uncertainty in activity data. First of all, we created a color palette (see Figure 7.5a) as many participants stated that colors would help them to quickly decide whether they reached their goal. We used a color scheme from red to green with green being the color for the reached goal. Second, multiple participants suggested to not show a single value, but a range of the acquired steps that includes a certain measurement error. Therefore, we developed a textual range representation (see Figure 7.5b).

In addition to these basic components, we developed three graphical overviews that can be used instead or in combination with the textual range representation. The first graphical overview is a bar chart with a grey part to indicate the potential error of measurement (see Figure 7.6a). The second graphical overview shows an error bar in addition to the bar chart (see Figure 7.6b). The third graphical overview is a speedometer with two needles (see Figure 7.6c), as many graphical overviews for activity data currently use speedometers or pie charts. All graphical overviews use the color palette introduced in Figure 7.5a to indicate how far users are from reaching their goal.



(a) Bar chart with range (b) Bar chart with error bars (c) Speedometer

Figure 7.6: Three different graphical overviews for activity tracking data developed on the basis of the comments of participants in the online survey.

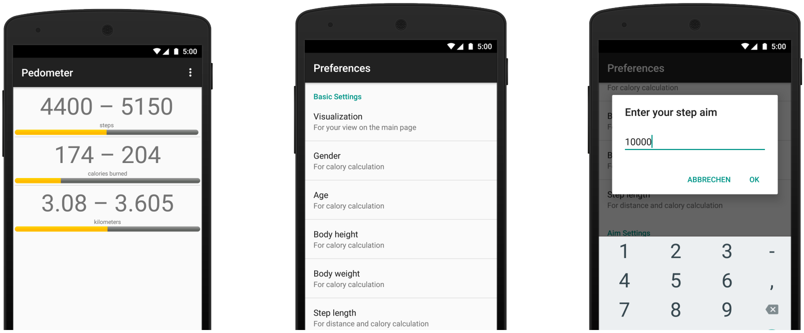
We conducted a user study with an Android application to explore whether users would understand and accept the new graphical overviews.

Android Application

We developed an Android application called *Pedometer* that tracks users' steps, and calculates calorie expenditure and the walked distance based on the step count. The application shows all values as ranges (see Figure 7.7a) instead of showing one single value. Users can enter personal details and their aim through the screen displayed in Figure 7.7b. They can adjust their step aim by simply changing a number (see Figure 7.7c). The application offers the three graphical overviews depicted in Figure 7.6, and users can choose which one they prefer as a default view. When clicking on a displayed range, the application shows their default graphical overview providing the possibility to swipe to see the alternatives.

Method

We invited the prospective participants to the lab to give them a short introduction to the study and the graphical overviews. Participants had to fill a pre-study questionnaire and install the application on their Android device. We also collected their contact details.



4400 – 5150
steps
174 – 204
calories burned
3.08 – 3.605
kilometers

Basic Settings
Visualization
For your view on the main page
Gender
For calory calculation
Age
For calory calculation
Body height
For calory calculation
Body weight
For calory calculation
Step length
For distance and calory calculation

Enter your step aim
10000

(a) Start screen of the *Pedometer* application

(b) Preferences in the *Pedometer* application

(c) Entering a goal in the *Pedometer* application

Figure 7.7: Screens of the *Pedometer* application.

We asked participants to use the *Pedometer* application for at least one week. Each day, we contacted participants twice to remind them to use the application and give them instructions for the day. On the first day, participants had to download any step tracking application of their choice from the Google Playstore and compare it to the *Pedometer* application. On the following three days, participants used one specific graphical overview of the *Pedometer* application each day to get an overview of all of them. On the fifth and sixth day, we asked participants to switch between the graphical overviews and compare them. On the seventh day, they were free to choose their favorite graphical overview to use on the last day of the study.

At the end of the week we again invited the participants to the lab, and asked them to answer a questionnaire about their opinion of the application and the graphical overviews.

Participants

We recruited 10 participants (6 male, 4 female) with an average age of 22 years, who used the *Pedometer* application for one week. In the pre-questionnaire all participants stated that they knew that measurement errors occur when using activity trackers. They estimated the error of measurements to lie between 7 %-10 %.

Results

In general, most participants liked using the *Pedometer* application. Seven out of ten stated that they would also like to continue using graphical overviews showing a range instead of a single value. However, they wanted this as an optional feature so that they could enable it depending on the sports they were doing. As they expected an activity tracker to track walking better than basketball they would enable the setting while playing basketball, but not while walking. Three participants also mentioned that the numbers of the range could be rounded to the next ten or hundred to make it easier to read. Four participants also explicitly stated that they liked the bar below the values which gave a rough and quick overview of the activity.

The three participants who did not like the range display and the novel graphical overviews stated two main reasons for their opinions. Two participants stated that two numbers are too complicated to read and do not allow for a quick overview. As they know about the errors of measurements, they perceive the single displayed value as an estimate anyway. Another participant stated that he was too used to the graphical overview he normally uses, and tried to calculate the average all the time.

At the end, we asked participants which one of the graphical overviews they preferred most. Six participants preferred the bar chart with the grey bar. They stated that it was easy to understand and to determine how far they still were from reaching their goal. Three participants preferred the speedometer. They liked its compactness and found it motivating. However, they suggested a display of colored triangle instead of two needles. One participant preferred the bar chart with the error bar as it contained the exact value and the measurement error in one visualization. However, this opinion was not shared by other participants.

Discussion

In general, participants liked to have graphical overviews including uncertainty, but wanted the feature to be optional. A graphical overview for activity data should therefore either allow users to switch between a single value and an interval or show both at the same time. The bar chart with the grey bar could for example also show a line for the measured value and just show the grey bar to raise awareness for the uncertainty. Additionally, multiple different graphical overviews could be offered for users to pick their favorite one. The speedometer, for example, was only preferred by three participants, but the others did not dislike it even though it was not their preferred option.

As participants complained about the range not being easy to grasp and remember, it could be interesting to show one color or a progress bar only instead of concrete numbers. This would make the information quickly graspable without using a concrete number that may imply wrong reliability. Users would only be able to roughly estimate the current value. This may be too drastic, but on the other hand would also preserve privacy as only the user knows the personal goal and therefore can interpret the color or progress bar correctly.

7.2.3 Implications

Graphical overviews for activity data should communicate the uncertainty of the data. On one hand, users prefer to get more information, but on the other hand want to be in control of the shown data. Therefore, graphical overviews should either include a concrete value and an additional indicator for uncertainty or make the display of the uncertain information optional. Nevertheless, showing the potential measurement errors can educate users not aware of them to recognize the associated uncertainty and take them into account for decision-making. However, manufacturers of activity trackers might not be eager to share information about measurement errors with their customers as devices could be compared by their reliability more easily. On the other hand, sharing such information could create more trust in products that communicate uncertainty.

One other important aspect is that a graphical overview for activity data still has to be easy to grasp. Adding information about the measurement error results in more visual clutter, which may add too much cognitive overload for users. It is therefore very important to keep the simplicity of the graphical overview when adding extra information.

7.3 Decision-making under Uncertainty

Previous work on uncertainty visualization focused on a small number of representations, variations of one representation, or a very specific task (e.g. finding the mean). Although it is known that showing uncertainty leads to better decisions, the relation between different degrees of uncertainty information included in a visualization and decision-making is still unclear. In contrast to prior work, our approach is to compare a large number of representations based on the amount of uncertainty information they include instead of merely comparing different visualization techniques.

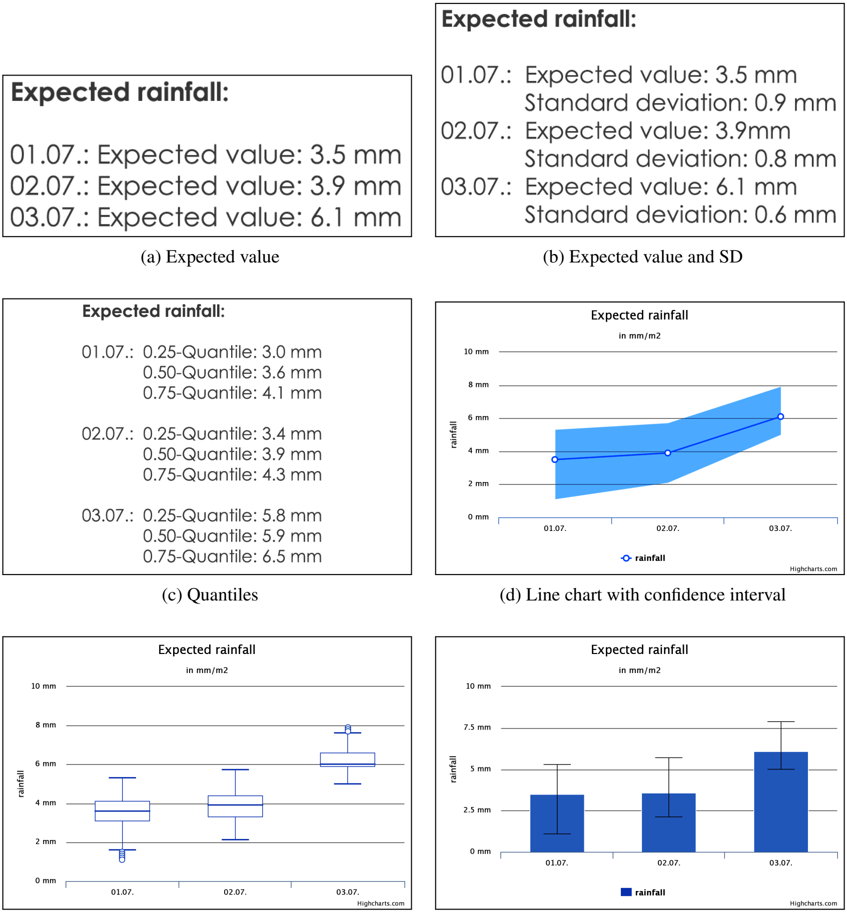
The main goal of the work presented in this section is to understand how the presented amount of uncertainty influences decision-making and whether other confounding factors besides the amount of uncertainty information play a role for users' preference. In the following, we classify 12 representations and compare them in an online survey. We further present the results of a follow-up experiment in which we compared four of the representations in an online Facebook game to understand their influence on decision-making.

7.3.1 Classification of Representations

Building upon prior research, we selected twelve representations (see Figure 7.8 and Figure 7.9) with different properties for communicating uncertainty information. These representations are classically used to communicate uncertain data to scientists and the general public. All representations show the expected rainfall for the next three days. Three representations use a textual representation of the information, whilst the other nine are graphical. We use a line chart, a box-and-whisker plot, bar charts, stacked bar charts, stacked area diagrams, shaded bars, and function graphs. Table 7.6 clusters the representations based on the degree of uncertainty information included in the representations. Representations can either contain no information about the uncertainty, aggregated information, detailed aggregated information, or fully detailed information. This classification based on the degree of uncertainty information included in a representation, is a novel way of classifying representations. We argue that this classification simplifies the comparison of representations including uncertainty information.

7.3.2 Online Evaluation

We first aimed at evaluating all representations to reduce their number for follow-up experiments. We therefore conducted an online survey to compare all twelve representations according to their perceived value for decision support and their easiness.



Expected rainfall:

01.07.: Expected value: 3.5 mm
02.07.: Expected value: 3.9 mm
03.07.: Expected value: 6.1 mm

(a) Expected value

Expected rainfall:

01.07.: Expected value: 3.5 mm
Standard deviation: 0.9 mm
02.07.: Expected value: 3.9mm
Standard deviation: 0.8 mm
03.07.: Expected value: 6.1 mm
Standard deviation: 0.6 mm

(b) Expected value and SD

Expected rainfall:

01.07.: 0.25-Quantile: 3.0 mm
0.50-Quantile: 3.6 mm
0.75-Quantile: 4.1 mm

02.07.: 0.25-Quantile: 3.4 mm
0.50-Quantile: 3.9 mm
0.75-Quantile: 4.3 mm

03.07.: 0.25-Quantile: 5.8 mm
0.50-Quantile: 5.9 mm
0.75-Quantile: 6.5 mm

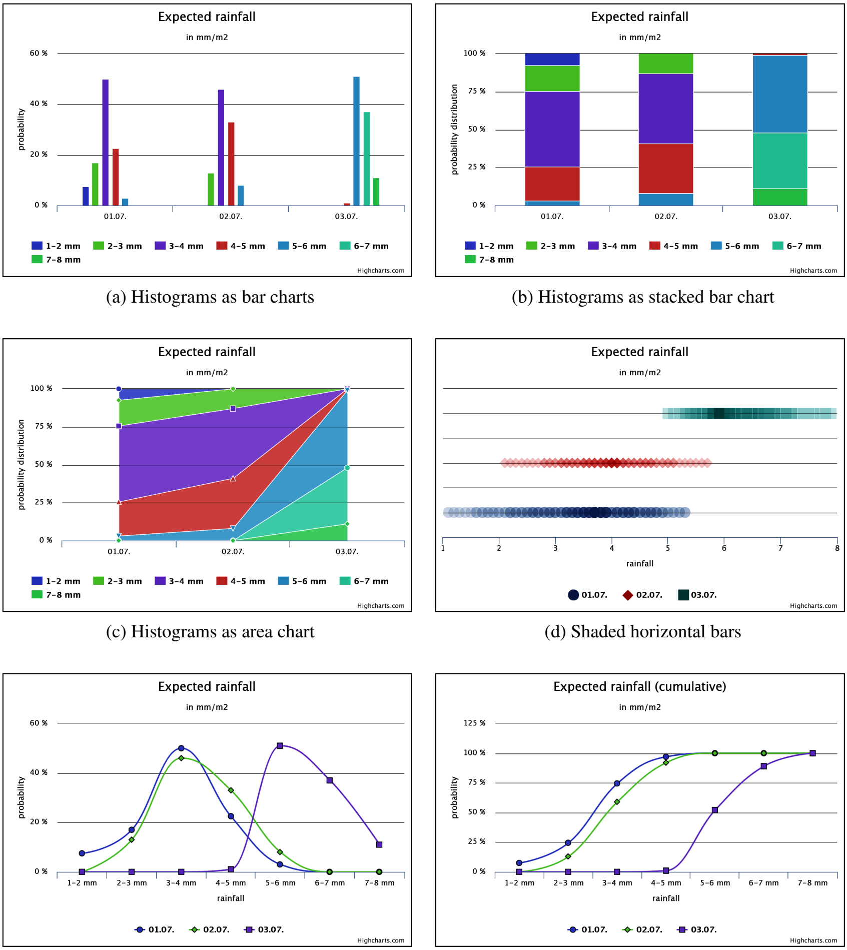
(c) Quantiles

(d) Line chart with confidence interval

(e) Box-and-whisker plot

(f) Bar chart with error bars

Figure 7.8: First half of the representations used in the online survey showing no or aggregated uncertainty information.



(a) Histograms as bar charts

(b) Histograms as stacked bar chart

(c) Histograms as area chart

(d) Shaded horizontal bars

(e) Probability distribution function plot

(f) Cumulative probability distribution function plot

Figure 7.9: Second half of the representations used in the online survey showing detailed aggregated or detailed uncertainty information.

Table 7.6: Degree of uncertainty information included in the representations.

	Textual Representation	**Graphical Representation**
No Uncertainty Information	REP1: Expected values	-
Aggregated Uncertainty Information	REP2: Expected values and standard deviation REP3: Quantiles	REP4: Expected values and confidence interval REP5: Quantiles REP6: Expected values and standard deviation
Detailed Aggregated Uncertainty Information	-	REP7-9: Aggregated probability density function REP10: Color-coded probability density function
Detailed Uncertainty Information	-	REP11: Probability density function REP12: Cumulative probability density function

Method

We first asked participants for demographic information such as age, gender, highest degree, and background. We then introduced the study with a scenario description. Participants should imagine that they are a farmer who wants to grow crops. The plants need a certain amount of rain to survive and grow. A weather forecast supports them to decide which crops to grow. As expected from a weather forecast, it is uncertain. After reading the scenario description, participants started the main part of the survey. Each of the following twelve pages of the survey contained one of the representations. The order of representations was randomized across participants to reduce sequence effects. For each representation, participants had to indicate their level of agreement on a five-point Likert scale from “totally disagree” to “totally agree” for the following four statements:

- The representation supports me in making a decision.
- I am familiar with the representation.
- The representation is easy to understand.
- The representation is visually appealing.

Participants

In total, 90 participants (36 female, 54 male) fully completed the online survey. Participants' age ranged from 18 to 82 years ($M = 31.0$, $SD = 12.6$). 45 % of our participants had a university degree, 28 % a high school diploma, and a further 20 % had vocational training. The other participants either had a minor or no degree at all. Participants had diverse backgrounds such as computer science, economics, teaching, mechanics, and services.

Results

We converted the Likert items to numbers; 1 corresponding to "totally disagree" and 5 corresponding to "totally agree". For each representation, we then calculated the mean for each of the four statements and an overall mean. For each statement, we also conducted a Friedman test and Wilcoxon signed rank tests with an applied Bonferroni correction. The Friedman test showed that there was a statistically significant difference between the twelve representations for each statement. The Wilcoxon signed-rank test revealed that representation 1, 4, 7, and 11 performed significantly better than the majority of other representations for at least one judgment each. Representation 3 was rated significantly worse than the majority of representations on three statements.

We additionally ran a Spearman's rank-order correlation to determine relationships between the statements and the degree of uncertainty information that the representations show. As expected, we found strong positive correlations between all pairs of Likert items, which all were statistically significant ($p < .001$). However, the correlation coefficients did not reveal any significant positive or negative correlation between the statements and the degree of uncertainty information of the representations. The only exception was one moderately positive correlation with the Likert items for visual appeal. We assume that this correlation occurred because the representations with low degrees of uncertainty information were textual representations and therefore less appealing for participants.

Discussion

The results of the online survey show that participants had different opinions on how much the representations would help them to make a decision. Surprisingly, the four best-rated representations taking the overall mean show a different degree of uncertainty information each. We found that participants do not rate the representations based on their degree of uncertainty. Instead, factors such as familiarity, easiness to understand, and visual appeal have a huge influence

Table 7.7: Calculated mean values for the level of agreement on a five-point Likert scale from totally disagree (1) to totally agree (5) with the statements: S1 - The representation supports me in making a decision., S2 - I am familiar with the representation., S3 - The representation is easy to understand., S4 - The representation is visually appealing., and O - the overall mean values.

Representations	**S1**	**S2**	**S3**	**S4**	**O**
1: Expected values	3.6	**4.3**	**4.3**	2.3	**3.6**
2: Expected values and standard deviation	3.7	3.9	3.3	2.1	3.3
3: Quantiles	2.9	2.8	2.4	1.7	2.4
4: Line chart with confidence interval	**4.1**	3.7	4.1	**4.0**	**4.0**
5: Boxplot	3.2	2.7	2.5	2.5	2.7
6: Bar chart with error bars	3.5	3.1	3.1	3.0	3.2
7: Histograms as bar charts	3.9	4.2	3.7	3.7	**3.9**
8: Histograms as stacked bar charts	3.5	3.5	3.2	3.5	3.4
9: Histograms as area chart	3.2	2.9	2.8	3.4	3.0
10: Shaded horizontal bars	3.7	2.3	3.6	3.6	3.3
11: Probability distribution function	3.9	3.9	3.5	3.6	**3.7**
12: Cum. probability distribution function	3.4	3.5	2.9	3.5	3.3

Table 7.8: Spearmen's rho for a Spearman's rank-order correlation between Likert scale items of the online survey and the degree of uncertainty information of the presented representations. Statistically significant values are marked with asterisk(s). $^{**}p < .01$, $^{*}p < .001$

	Decision Support	**Familiarity**	**Easiness to Understand**	**Visual Appeal**
Degree of Uncertainty	0.048	-0.040	-0.084^{**}	0.365^{*}
Decision Support	-	0.505^{*}	0.670^{*}	0.527^{*}
Familiarity	-	-	0.609^{*}	0.530^{*}
Easiness to Understand	-	-	-	0.530^{*}

on whether participants found a representation suitable for decision making. Our work indicates that these confounding factors have to be considered when displaying uncertain data.

Based on our work, the most promising candidates for uncertainty visualization with different degrees of uncertainty can be selected for future studies with the general public.

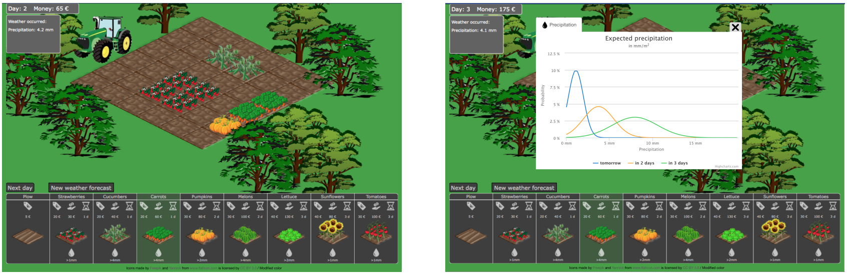
7.3.3 Evaluation in the Wild

We developed a web-based game called “Farm Smart” where we included the four representations that performed best in the online evaluation. We ran a pilot study with 12 participants and based on their feedback improved the game before publishing it on Facebook.

Method

In the turn-based Facebook game “Farm Smart” (see Figure 7.10), players can buy, plant, and harvest crops to earn as much money as possible. Each crop needs a certain amount of precipitation and has a certain ripening time between one and three days. In order to successfully grow, the precipitation value of a crop has to be fulfilled during the whole ripening time. To decide which crop to plant, players can look at the forecasted precipitation of the next three days by clicking on a button (see Figure 7.10b). After planting as many different crops as desired, a player ends the day by clicking on the “Next day” button. The weather (displayed in the upper left corner of the screen) and specific field icons indicate whether the crop survived or withered (see Figure 7.10a). Ten days (i.e. turns) correspond to one game. We included the four representations performing best in the online survey for displaying the weather forecast. The representation would change during a game. Each player could in total play four games; one with each of the four representations and randomly assigned to one of the 24 permutations of the order of the representations at the first start of “Farm Smart”. To recruit players, we shared the game link on Facebook and posted advertisements in online gaming communities.

Requirements and Prices of Crops. The seed costs, sale prices, ripening times, and weather requirements of crops were specified differently to ensure a wide range of possible decisions in all weather situations. Players are motivated to equally consider more demanding crops, as higher requirements or ripening times



(a) General view of “Farm Smart” with the possibility to plant crops on the field displayed in the middle of the screen.

(b) Weather forecast as displayed in “Farm Smart”.

Figure 7.10: Screenshots of our Facebook game “Farm Smart” to compare representations including a different degree of uncertainty information.

are rewarded with a higher money gain when selling the crops. The rewards for the overall game were chosen to be pareto-optimal.

Weather Calculation. The weather forecast in the game is based on real weather measurements of a meteorological station[5] in Stuttgart. The values used in the game cover a continuous period of 62 days between spring and summer 2001. This period offers diverse weather conditions and is in general rich in rainfalls. For every game, the start day is uniformly selected at random from the first 49 days of the period to ensure that the forecasts are available for the whole game. A weather forecast in “Farm Smart” corresponds to a set of three modified Gaussian distributions. For a day d the forecast for the n-th day ($n \in \{1,2,3\}$) is constructed as follows:

1. The expected value μ' of the Gaussian distribution is calculated by offsetting the real value of the day $d+n$ by a factor $n \cdot \alpha \cdot X$ with $\alpha > 0$ and a sample $X \sim \mathcal{N}(0,1)$. This corresponds to the assumption that the magnitude of the inaccuracy in weather forecasts is approximately normally distributed and increases linearly with the temporal distance to the predicted event. As a negative μ' for the precipitation cannot be reasonably interpreted in our context, μ' is set to 0 if $\mu' < 0$.
2. The standard deviation of the Gaussian distribution is chosen as $n \cdot \sigma$ with $\sigma \in \mathbb{R}_0^+$. This corresponds to a horizontal compression of the distribution's

[5] Source: German Weather Service, `http://www.dwd.de`

graph and represents the linearly increasing uncertainty within temporal distant weather forecasts. It follows the Gaussian distribution $\mathcal{N}(\mu', n \cdot \sigma)$ or in more detail $\mathcal{N}(\text{RealValue}(d+n) + n \cdot \alpha \cdot X, n \cdot \sigma)$.

We tested and compared different parameters α and σ before and in our pilot study. We then chose $\alpha = 0.8$ und $\sigma = 1.0$, because the values conveyed a suitable amount of uncertainty in the prediction.

Representations. We decided to use the four representations that performed best in the online evaluation and implemented them using HighCharts[6]. We shortly explain these visualizations in more detail:

Text: It shows the expected value, which gives no information at all about the uncertainty (see Figure 7.11a).

Line chart: It also shows the expected value, but adds quantiles, which were chosen as the 0.05 and the 0.95 quantile (see Figure 7.11b).

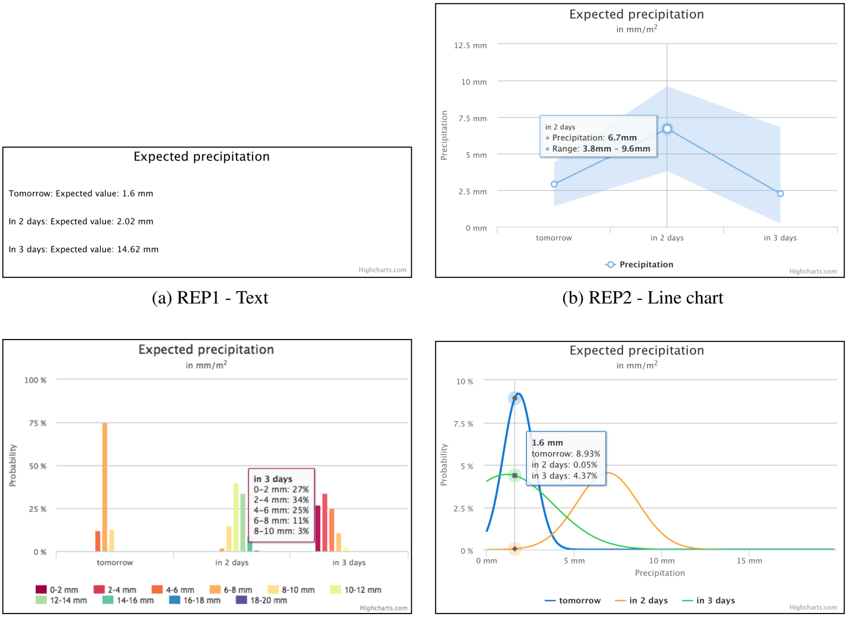
Bar chart: It visualizes the sum of probabilities within a certain range (see Figure 7.11c). For the bar colors, we used the HCL color space to achieve optimal differentiability and provoke equally intense perceptions.

Probability distribution function: It displays the underlying distributions in full detail in a probability distribution function graph (see Figure 7.11d).

The representations show an increasing amount of information about the uncertainty: the first representation shows no information, the second representation shows aggregated uncertainty information, the third representation shows aggregated detailed information, and the fourth representation shows all details.

Data Collection. We collected two types of data. First, we collected survey data, which included personal data and subjective feedback about the representations. After finishing a game, we asked players to rate the used representation on a five-point Likert scale with the statements used in the online survey (see Subsection 7.3.2). Second, we logged all game parameters such as accumulated money, representation type, forecasted weather, and the occurred weather. We additionally counted on how often participants clicked the button to open the weather forecast.

[6] HighCharts, `http://www.highcharts.com`



(a) REP1 - Text

(b) REP2 - Line chart

(c) REP3 - Bar chart

(d) REP4 - Probability distribution function

Figure 7.11: The four representations each display weather forecasts with a different amount of uncertainty information.

Results

We analyzed the data of 44 players who in total played 98 games (on average 2.2 games per player, $SD = 1.3$) composed of 991 turns. We only considered games in which player actively opened the weather forecast. On average, participants played 247.75 turns ($SD = 34.1$) per representation. As not all players played four games, they might only have experienced a subset of representations.

Survey Data. In total, 38 players completed the short survey (29 male, 9 female) after completing 88 games. We grouped their answers into three categories: (1) agree (includes agree and strongly agree), (2) neutral, and (3) disagree (includes disagree and strongly disagree). As Table 7.9 shows, players preferred the line chart and the bar chart.

Log Data. We calculated three metrics to explore the two questions: (1) How did the representations affect risk taking? and (2) How did the representations

Table 7.9: Percentage of agreeing players in the survey. Green shows the highest values. Red shows the lowest values.

	Survey Qustion (after each game)	**Percentage (%) of Aggrement**			
	Agreement = Agree, Strongly Agree	**REP1**	**REP2**	**REP3**	**REP4**
1	Familiar	59.1	66.7	61.9	47.6
2	Easy to understand	54.6	66.7	66.7	38.1
3	Visually appealing	22.7	85.3	33.3	38.1
4	Supports decision-making	31.8	54.2	57.1	28.6

Table 7.10: Overview of log data collected in the game.

	Average		**% of turns ending**		**Average money**	
REP	**weighted risk**		**with winning/losing**		**gained/lost**	
	Mean	**SD**	**Win**	**Lose**	**∅ gain**	**∅ loss**
REP1	12.8	24.9	24.1 %	47.4 %	435.3	173.7
REP2	18.7	31.5	17.8 %	54.4 %	448.1	130.2
REP3	17.3	29.1	23.5 %	49.8 %	772.2	170.3
REP4	20.3	33.4	15.8 %	49.0 %	612.4	161.2

affect decision-making? Table 7.10 shows a summary of the results. The metric "mean weighted risk" calculates the average risk per turn players were willing to take. The average indicates that players took most risk when using the probability function and least risk with the textual representation. For the decision-making, we looked at the "percentage of won turns" and the "mean positive money gain". Players won most turns when using the textual representation, however, they on average won more money when using the bar chart.

Discussion

The textual representation without uncertainty information led to players taking the lowest risks, but also resulted in the lowest amount of gained money and the highest amount of lost money. For the line chart and the bar chart, players took roughly the same amount of risk which is supported by the survey as both representations were preferred. Nevertheless, players won more turns using the bar chart and on average gained the highest amount of money. The line chart on the other hand resulted in the highest losses. The probability distribution function led to the most risky decisions, but therefore also a higher percentage of lost turns. The survey supports this as players seemed to rate it as a complex representation.

On the other hand, players who truly understood the representation were able to perform very well, which can be seen from the high average of the money gain.

7.3.4 Implications

The survey showed that more uncertainty information is not necessarily better or perceived as better for decision-making by people. Familiarity, easiness to understand, and visual appeal are some factors that have to be taken into account when designing representations for uncertainty communication.

The results of the experiment indicate that people do not favor representations without uncertainty information, but these representations are a “Low risk, low reward” option as they maintain high winning rates, but with low average of won money. Representations with detailed uncertainty information such as the probability distribution function are a “High risk, high reward” option, as they are associated with highest winning amounts when understood correctly. However, they might also cause high losses. Representations with aggregated uncertainty information were perceived most helpful and understandably by participants. However, it encouraged them to take incalculable risks. It seems that such representations lead to wrong assumptions and overestimation of the amount of included information. Finally, representations with aggregated detailed uncertainty information seem to offer a good compromise between understandability, taking educated risks, and achieving to win with high gains.

7.4 Insights for Communicating Uncertainty

In this chapter, we developed and explored output methods for uncertainty in interactive systems. We focused on the current state of uncertainty communication in mobile applications, users’ preferences for the future of uncertainty communication, and how different degrees of uncertainty information included in representations influence decision-making.

Currently, uncertainty is seldom communicated. If it is communicated, vague linguistic expressions are used. However, users raised a preference for more information about uncertainty depending on the context of an application. Nevertheless, a compromise between the preference for more information and the added cognitive load needs to be found. The main reason is that users still demand the information to be easily graspable.

Overall, we found that in terms of decision-making, it is not necessarily always better to show uncertainty information. The choice of representation should take into account confounding factors such as familiarity, visual appeal, and easiness to understand. Our classification of how much uncertainty information is included in a representation can additionally help to detect representations that might give users a wrong feeling of control and understanding. Based on our findings, we recommend to use representations including aggregated details uncertainty information such as a histogram or dot plot.

In the next chapter, we focus on the third important aspect of how users deal with uncertainty in interactive systems: the interpretation of uncertain data. We explore how humans make predictions, how interactive systems can support the aggregation of uncertain data, and the internal models humans use for making sense of conflicting uncertain information.

Chapter 8

Interpretation

We explored humans' ability to interpret uncertain data by conducting three individual research probes. In the first probe, we examined how good humans are at making predictions themselves and how that influences their personal understanding. In the second probe, we explored different possibilities to support humans in comparing uncertain data from multiple sources by using different aggregation mechanisms. In the third probe, also related to conflicting data, we aimed to understand humans' internal models of making sense of conflicting data from different sources with a different degree of uncertainty information provided. All probes were evaluated in user studies.

The main goal of this chapter is to better understand how humans interpret uncertain data and how they can be supported to correctly interpret uncertain data.

Parts of this chapter are based on the following publications:

- M. Greis, E. Avci, A. Schmidt, and T. Machulla. 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*, pages 828–840, 2017.
- M. Greis, T. Dingler, A. Schmidt, and C. Schmandt. Leveraging User-made Predictions to Help Understand Personal Behavior Patterns. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17*, pages 104:1-104:8, 2017.

Parts of this chapter are also planned to be published as follows:

- M. Greis, A. Joshi, A. Schmidt, and Tonja Machulla. Uncertainty Visualization Improves Humans' Choice of Internal Models for Information Aggregation.

8.1 Humans Predictions

Smartphone applications and wearable devices count how many steps we take, how often we unlock our phone, or how many words we read. Collecting such data leads to large datasets that are presented to people in order to motivate behavior change. This is semi-successful and prone to only have short-term effects [Ledger and McCaffrey, 2014]. One reason for this is the abstract nature of the data which on its own is insufficient to make users think about their behavior patterns [Choe et al., 2014].

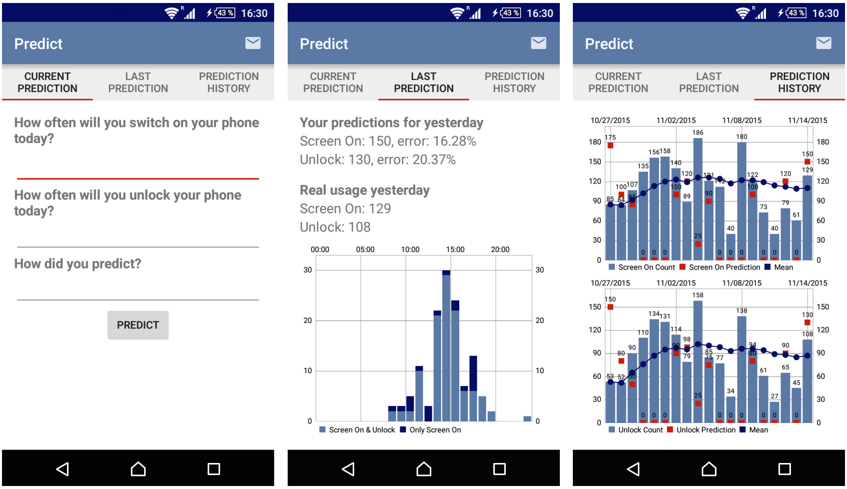
The main goal of the work presented in this section is to understand whether user-made predictions can serve as a tool to improve users' reasoning about predictions and their understanding of their own behavior patterns. To make personal predictions, users probably have to think about their behavior more carefully and their prediction is prone to including uncertainty as they cannot exactly know their behavior beforehand. We were interested in how people adapt their predictions and whether they improve their predictions over time. We decided to focus on mobile phone usage as this is easy to track without using additional sensors than the smartphone. Additionally, current research suggests that the average usage of applications is below one minute [Böhmer et al., 2011] and that checking habits lead to longer usage times [Oulasvirta et al., 2012]. We therefore built an Android application called *Predict* that tracks the number of users' screen-ons and unlocks. Users can additionally predict their behavior for the current day.

8.1.1 Android Application

We developed the Android application *Predict* that allows users to predict their mobile phone screen-on and unlock counts (see Figure 8.1). Each day on their first mobile phone usage, users received a notification that asked them to predict how often they would turn on their screen and unlock their phones on this current day (see Figure 8.1a). As soon as they entered their prediction, the application showed the predicted values instead as a simple text. As support for their predictions, users had detailed statistics about their usage from the last day (see Figure 8.1b) and a history diagram of their real usage and predicted values (see Figure 8.1c). To extract the usage behavior, the application counted the Android events SCREEN_ON, SCREEN_OFF, and USER_PRESENT. The application did not send any data to an external server unless participants pressed the button in the top right corner of the application. This functionality was implemented to increase participants' trust in the application.

8.1.2 Method

We invited prospective participants to install the application and use it for 14 consecutive days. After this usage period, we sent them a reminder to share their data with us and provided an online questionnaire. The questionnaire consisted of two parts. In the first part, participants had to indicate their level of agreement with ten statements (see Table 8.1) on a five-point Likert scale. In the second optional part, we asked for qualitative feedback on different aspects such as when, why and how they predicted, how they felt about making predictions, whether the predictions influenced their usage behavior, and whether they learned something during the course of the two weeks.



How often will you switch on your phone today?

How often will you unlock your phone today?

How did you predict?

Your predictions for yesterday
Screen On: 150, error: 16.28%
Unlock: 130, error: 20.37%

Real usage yesterday
Screen On: 129
Unlock: 108

(a) Making a prediction (b) History for the last day (c) Complete usage and prediction history

Figure 8.1: Screenshots of the *Predict* application that show the three main screens of the application.

Table 8.1: Statements presented in the online questionnaire.

ID	Statement
S1	I liked to make prediction daily.
S2	I always looked forward to see how good my prediction was.
S3	I wanted to improve my prediction every day.
S4	I looked at the historic data before making a new prediction.
S5	I always used the same strategy for making my prediction.
S6	I think that my predictions improved over time.
S7	Making the predictions influenced my usage behavior.
S8	I tried to use my mobile device less.
S9	I will continue to make predictions with the app.
S10	I will continue to look at the historic data with the app.

8.1.3 Participants

Twelve participants (10 male, 2 female) with an average age of 27.0 ($SD = 11.7$) installed our application. Eight were students and the other four wage earners. All used the application for at least 14 days, and most continued voluntarily for at least a few more days.

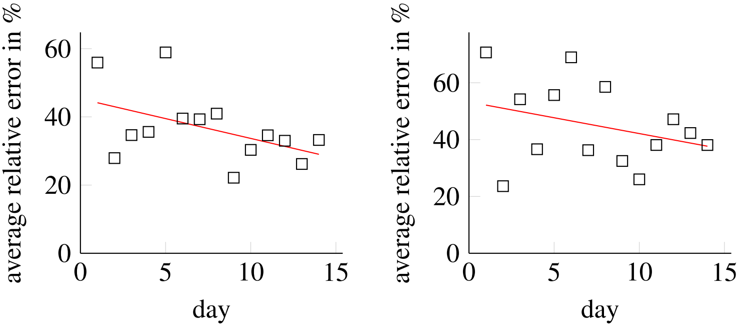
8.1.4 Results

We registered a total of 9,317 unlock events, 6,576 additional screen-on events (no unlock performed), and 336 predictions. In general, the number of screen-ons and unlocks varied highly across participants. They unlocked their phone between 2 and 264 times per day ($M = 55.8$, $SD = 41.7$). In addition, they turned on their screen between 2 and 399 times per day ($M = 39.4$, $SD = 52.3$). In the following, we outline participants' prediction accuracies and behavior over time. For prediction accuracy, we calculated the absolute value of the relative error in percent.

Screen-on & Unlock Predictions. For the screen-on predictions, participants had an average relative error of 36.7 % ($SD = 34.1\,\%$). For the unlock prediction, the relative error was even higher with 44.9 % ($SD = 50.4\,\%$). Participants' predictions, however, improved over time. Comparing the error rate of the first day and the last day of the study, it decreased for more than 20 % for screen-on predictions ($M = 33.2\,\%$, $SD = 31.3\,\%$) and more than 30 % ($M = 38.1\,\%$, $SD = 37.6\,\%$) for unlock predictions. Figure 8.2 shows the development of the average relative errors over time and the resulting regression lines.

Online Questionnaire. We analyzed participants' Likert scale ratings and provide the results in Figure 8.3. For the exact statements S1 to S10, see Table 8.1. To better present the ratings of participants, we converted the ratings to numbers from 1 for "totally disagree" to 5 for "totally agree".

Half of the participants liked to make predictions daily, while most of the other participants were neutral ($S1$, $M = 3.3$, $SD = 1.0$). Participants who liked to predict their behavior found it interesting and challenging: "*At first I thought it would be annoying, but then I was surprised that it really was interesting and somehow challenging to better my predictions.*" (Male, 22 years). Neutral participants did not associate any feelings with it: "*No special feeling. It was an ordinary task such as to set the alarm.*" (Male, 28 years). Two participants did



(a) Regression line for predictions of screen-ons. (b) Regression line for predictions of unlocks.

Figure 8.2: Regression lines for the average relative error of screen-on and unlock predictions.

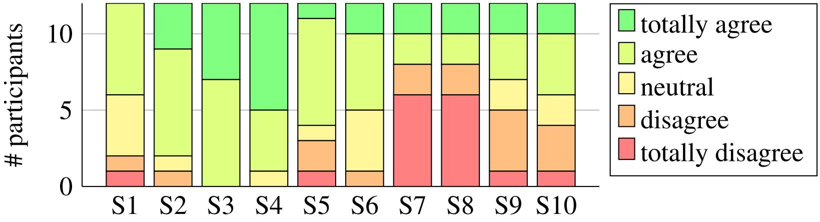


Figure 8.3: Results of the online survey showing the agreements of 12 participants on a five-point Likert scale with the statements outlined in Table 8.1.

not like to make predictions: "*It got annoying over time. That I wasn't accurate at all didn't help.*" (Male, 23 years).

All besides two participants looked forward to see how good their prediction was ($S2$, $M = 4.0$, $SD = 0.9$). Participants were also eager to improve their predictions from day to day ($S3$, $M = 4.4$, $SD = 0.5$), however, only seven participants had the feeling that their predictions actually improved ($S6$, $M = 3.7$, $SD = 0.9$).

We further looked into participants' prediction strategies. Most of the times, participants used the historic data to make their prediction ($S4$, $M = 4.5$, $SD = 0.7$) and did not change their prediction strategy ($S5$, $M = 3.4$, $SD = 1.2$). Participants started their predictions rather randomly and then developed different strategies, for example focusing on the average: "*I started with pretty random predictions for the first few days. Then as I saw a kind of average I focused on that. I guess after a week or so I thought about the day and what is going to happen. I also always checked for the last day's accuracy because I thought of it as my average usage.*" (Male, 21 years), recognizing patterns: "*I somehow tried to weigh different factors. First of all I made predictions using the last day. After I collected some data, I also used some patterns I could recognize (e.g. specific day of the week or activities).*" (Male, 20 years), or identifying influencing factors: "*At the beginning I used yesterday's historic data and prediction to make today's prediction. But that helped only part way because my phone use depends on many factors - day of week, whether in Boston or travelling, whether I walk or bike to work, even whether it is raining (don't want the phone to get wet).*" (Male, 63 years).

We asked participants to outline when they predicted and found that three of them did so in the morning when getting up or unplugging their phone from the charger: "*Morning, when I unplugged the phone. Often just before I leave the house.*" (Male, 63 years). Five participants predicted directly after midnight while three more specified that they either predicted at midnight or in the morning depending on how long they stayed awake. Six explicitly referred to predicting when receiving the notification: "*I always predicted before I went to sleep somewhat after midnight. This was the first time the app reminded me to do so.*" (Male, 25 years).

Participants perceived the influence of the application very differently ($S7$, $M = 2.3$, $SD = 1.7$). One participant, for example, stated that he forgot about the application until the next day: "*I didn't recognize any influence... after making the prediction I usually forgot about the app till the next day.*" (Male, 23 years). The four participants that agreed with the statement also agreed that they tried to use their phone less ($S8$, $M = 2.3$, $SD = 1.7$): "*I tried to look less often at my phone. I have a notification light, so sometimes in the past I would still turn the screen on to double check if nothing is on. I wouldn't do that anymore because for 'predict' it would be counted. I also tried not to unlock it that often anymore or not unlock it randomly without having a real purpose apart from being bored. In general, I would say it supported me very much in being more aware of my phone usage as finally I had numbers that would back up how often I am using it.*" (Female, 30 years).

Participants learned different things while using the application. One participant stated that he did not learn anything: "*My behavior was quite random so I often* *predicted totally wrong.*" (Male, 25 years). Two participants realized that they used their phone more often than they expected: "*That I looked at my device more than I thought I did.*" (Female, 28 years). Five participants learned and recognized patterns on where, when, and why they used their phone: "*I use the phone a lot less on the weekend, I hadn't quite realized how much less. How I travel makes a big difference on phone use. I don't turn it on as much as I thought I might.*" (Male, 63 years).

We also asked participants whether they would continue making predictions with the application ($S9$, $M = 3.1$, $SD = 1.3$) and whether they would continue to look at the historic data of the application ($S9$, $M = 3.3$, $SD = 1.3$).

8.1.5 Discussion & Implications

The results of our study show that participants' predictions improved over time. Participants tried to build their own internal model of their usage behavior and trained this model over time. They used strategies such as focusing on the average, recognizing patterns, or recognizing external factors to predict their behavior. Such strategies are also used for predictive algorithms and might be one reason for uncertain information. The capabilities of humans to make predictions could be used to support the explanation of predictive algorithms and increase users' understanding of uncertainty. The learning effect of making predictions could also be used to give users more trust in predictive algorithms. Participants, for example, learned that they used their phones much more than they thought before the study. A predictive algorithm predicting such high numbers with any proof or involved learning might not be seen as trustworthy.

Participants liked to make predictions and wanted to improve them. The internal model that they built for predictions also influenced their behavior. User-made predictions could therefore also be interesting to use in the context of behavior change. The eagerness of users to predict the right number and to stay in the range of the prediction could support such changes.

8.2 Aggregating Forecasts from Multiple Sources

For many short-term forecasts, users can choose between a large number of providers. Such forecasts may differ because providers use different models to create their forecasts. The model uncertainty, however, is seldom presented to the end-user. In our diary study (see Subsection 4.3.1), participants stated that they use multiple sources if they do not trust a single source. For example, before going on a hike, people will consult several weather forecast providers as facing a thunderstorm in the mountains can be deadly. Comparing multiple forecasts is mostly tedious and cumbersome as users may open several websites in different tabs, then try to remember the values of one website to compare it to another. High-level applications that support easy comparison have recently entered the market[7], but are not heavily in use yet. So far, there is no theoretical underpinning of how to design for easy comparison of weather forecasts, uncertain data in general, and the impact of different designs.

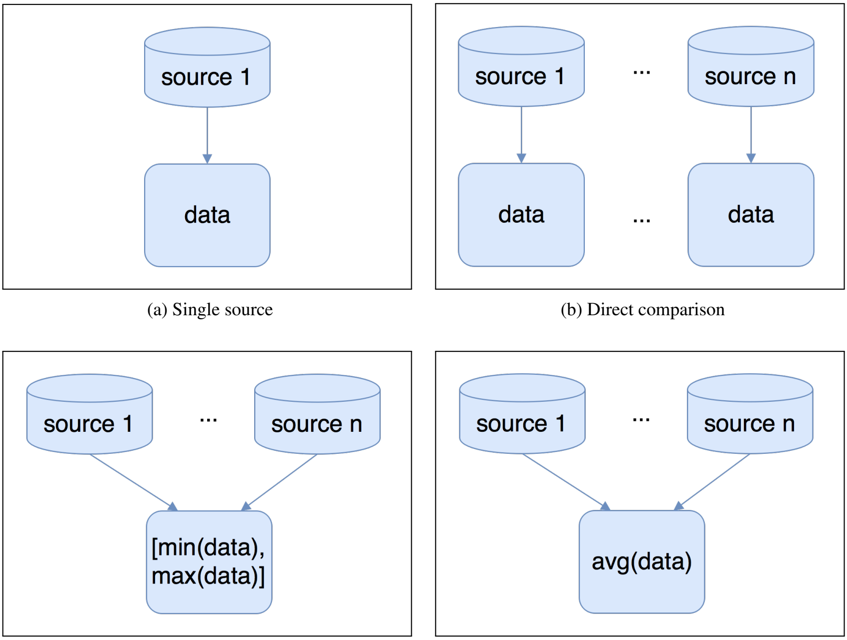
The main goal of our work is to understand how to design interfaces that support the easy comparison of uncertain data from multiple sources. We mainly aim to understand which designs are useful for users and increase their confidence in the depicted data. Therefore we compare three aggregation mechanisms: two mechanisms that computationally aggregate data and one that supports direct comparison without switching between tabs or applications. Computational and manual aggregation are likely to differ in terms of mental workload and the perceived amount of control.

In the following, we present the design rationale for the aggregation mechanisms, our detailed research questions, an online evaluation of the mechanisms, and an evaluation in-the-wild. In a hallway questionnaire, we asked six participants how many forecasts from different sources would be optimal to compare. One participant found two sources optimal, the other five participants found three sources optimal. We therefore decided to use three sources for all our designs.

8.2.1 Design Rationale for Aggregation Mechanisms

To reduce the cumbersome switching between different forecast providers and mental workload when comparing forecasts, two approaches can be followed.

[7] Climendo: http://climendo.com/, WeatherXM: http://weatherxm.exm.gr/



(a) Single source

(b) Direct comparison

(c) Range aggregation

(d) Mean aggregation

Figure 8.4: Many interactive systems, for example weather applications, show uncertain data from a single source (see a). To encourage the design of interactive applications that show uncertain data from multiple sources, we identified three aggregation mechanisms. The direct comparison allows users to look at data from multiple sources at the same time introducing mental workload to aggregate the data. The range and mean aggregation provide computationally aggregated data, which reduces the mental workload for the user.

First, the manual comparison could be supported by showing forecasts from different sources next to each other in a comparable format. Second, the computer could aggregate data from multiple sources which would reduce the workload further. The second approach may, however, reduce users' feeling of control as they do not necessarily follow decision aids and concrete advice [Joslyn and LeClerc, 2012]. In the following, we describe the aggregation mechanisms and the baseline condition in more detail (for an overview see Figure 8.4).

Single Source

A single source forecast (see Figure 8.4a) corresponds to how weather forecasts are currently displayed in most current weather applications. One weather provider shows exactly one forecast. This mechanism serves as a baseline condition for our study. It introduces the highest workload to users when it comes to comparing forecasts as they have to open different sources (e.g. websites or weather applications) and potentially remember the values if they cannot be displayed next to each other (e.g. due to small screen size on a mobile phone).

Direct Comparison

Direct or manual comparison (see Figure 8.4b) corresponds to showing multiple single source forecasts next to each other in one system. Thus, the forecasts can be perceived at one glance and no switching between different providers is needed. This is similar to the approach used by Frick and Hegg [2011]. The mechanism reduces the workload for the users as they do not have to remember the values and saves the time and effort to find different providers. However, users still have to manually compare and interpret the forecasts. Depending on the screen size, there could be constraints on how many sources an application can show on the screen without scrolling.

Range Aggregation

Range aggregation (see Figure 8.4c) computationally aggregates forecasts of multiple providers, but still enables users to make some decisions on their own. Morss et al. [2010] already suggested using a range representation for displaying uncertainty in weather forecasts. In contrast to their work, we do not use values from the same weather provider but show the minimum and maximum values from across multiple sources. This makes it easy to spot how much the values vary between providers without having to compare all single values manually. However, the mechanism would be affected significantly by one erroneous source.

Mean Aggregation

Mean aggregation (see Figure 8.4d) is identical to a single source forecast except that a mean of multiple sources is provided. The mechanism aggregates multiple values into a single value and therefore keeps the simplicity of a single source forecast, nevertheless it takes into account more information. Depending on the forecast information, a weighted mean based on the accuracy of sources could be used. If a mean is used it must be clearly communicated to users as they may

Table 8.2: Detailed research questions about the design for comparison of uncertain data.

RQ	Research Question
RQ5.1	Does the aggregation of multiple forecasts change the users' confidence in uncertain data?
RQ5.2	Do people prefer aggregated forecasts to single source forecasts?
RQ5.3	Do people prefer different aggregation mechanisms (manual vs. computation aggregation) according to the importance of a scenario?
RQ5.4	Do people prefer different aggregation mechanisms depending on the type of visual or textual representation?
RQ5.5	Can the theoretical findings be transferred to real world application usage?

otherwise interpret it as a single source forecast. Mean aggregation reduces the workload, however users might feel a loss of control as they do not have access to the values of different forecasts.

8.2.2 Detailed Research Questions and Hypotheses

Our main goal is to understand how to design for the comparison of uncertain data and how this influences users. We broke this goal down to five detailed research questions (see Table 8.2).

For RQ5.1, we assume that aggregation increases users' confidence in the presented data as users will get a better understanding for the uncertainty by comparing the sources. For RQ5.2, we assume that users generally prefer aggregated forecasts to single sources forecasts as aggregation adds information and therefore leads to more informed decisions. We assume that for RQ5.3, the preference of aggregation mechanisms depends on the importance of the scenario and context of users. While in important scenarios users may want more control, they will happily use aggregated forecasts to reduce workload in less important scenarios. We additionally assume that for RQ5.4, different representations influence the preference for aggregation mechanisms as representations may harmonize better with one or the other aggregation mechanism. RQ5.5 is a more explorative research question, but we assume that the theoretical findings will be transferable to the real world.

8.2.3 Online Evaluation

We first evaluated the three aggregation mechanisms and the baseline in an online survey where we aimed to understand users' preferences, the influence of the mechanisms on users' confidence, and mechanisms' relationship to different representations and scenarios of varying importance.

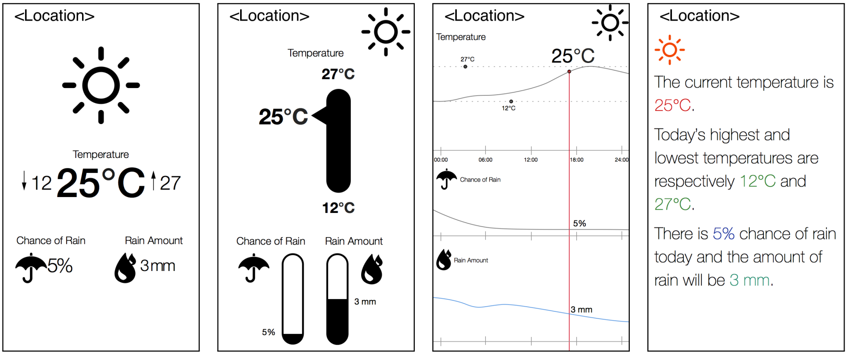
Design

We used a 4 x 4 x 3 within-subject design with three independent variables: *Aggregation mechanism* (with the four levels single source, direct comparison, range, mean), *representation* (with the four levels text, pictogram, bar, line), and *scenario* (with the three levels daily dress code, outdoor BBQ party, outdoor wedding). We measured participants' general confidence and their preference regarding all scenarios on a seven-point Likert item for each of the 16 combinations of aggregation mechanism and representation. For an explanation of the aggregation mechanisms, see Subsection 8.2.1.

Representations. As current weather forecasts use different representations, we also included different representations in our survey. To identify suitable representations, we investigated weather applications and websites, which mostly use pictograms with numbers or line charts. We added a text and a bar representation for more variety. Although the uncertainty in our study is not stochastic, we treat it as model uncertainty and use the ambiguation methods developed by Olston and Mackinlay [2002] for the range aggregation in bar and line charts. All representations show the location of the weather station, a pictogram of the current weather, and numerical values for the temperature (current temperature, daily minimum, daily maximum), the chance of rain, and the amount of rain. Figure 8.5 depicts the four representations for the single source condition.

Scenarios. We assembled a list of twelve different weather-related scenarios: (1) What to wear when going to work, (2) Whether to take the train or bike to work, (3) Outdoor soccer viewing event with friends, (4) Camping trip, (5) Organizing an outdoor wedding, (6) Packing for a trip, (7) Doing outdoor sports (e.g. swimming, hiking), (8) Planning an outdoor barbecue party, (9) Harvesting a crop, (10) Gardening, (11) Whale watching, and (12) Attending an outdoor concert.

To eliminate the time factor, we formulated all of the scenarios to take place "tomorrow". In a short hallway questionnaire with 15 participants, we determined the importance of accurate prediction for the scenarios. Participants had to



The current temperature is 25°C.

Today's highest and lowest temperatures are respectively 12°C and 27°C.

There is 5% chance of rain today and the amount of rain will be 3 mm.

(a) Pictogram (b) Bar graph (c) Line graph (d) Text

Figure 8.5: Exemplary sketches for the representation methods used in the online survey displaying a single source or mean aggregation forecast.

quickly decide for each scenario whether it was “casual important”, “somewhat important”, or “very important”. Based on the results, we selected the following three scenarios with varying importance:

Least importance (10 participants chose “casual important”): Daily Dress Choice - You are going to work/school tomorrow morning on a regular day and you need to decide what to wear.

Medium importance (11 participants chose “somewhat important”): Outdoor BBQ Party - You are planning a BBQ party for tomorrow and you want to know whether the weather will be good for having a party outdoors.

Highest importance (13 participants chose “very important”): Outdoor Wedding - You have your wedding tomorrow and it is an outdoor wedding. You need to decide whether you should make changes in the organization (like arranging a tent, or moving the wedding indoors).

Method

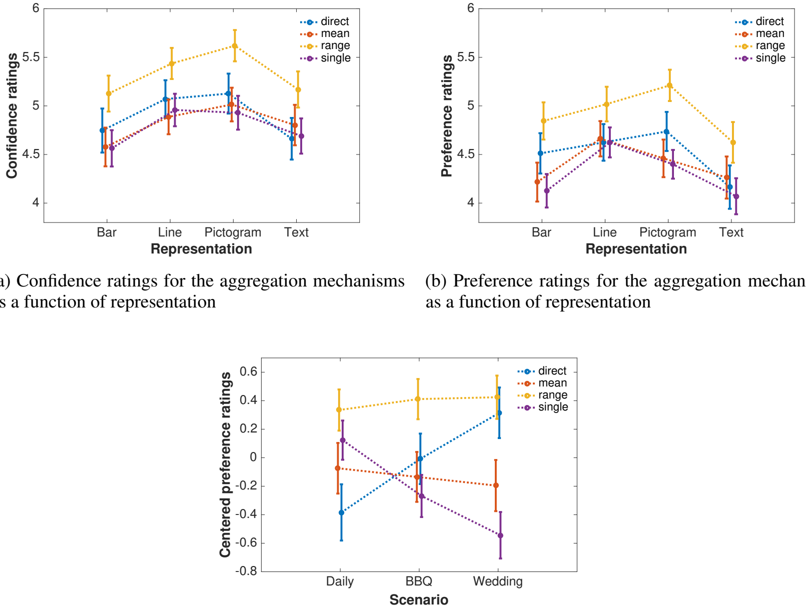
At the beginning of the survey, we collected demographic data and details about the habitual usage of weather forecasts. Participants then navigated through four pages; one for each representation. The order of representations was randomized across participants to reduce sequence effects. On each page, participants encountered an explanation of the representation, all aggregation methods and all scenarios. Additionally, each page contained four sketches, which showed the representation in combination with each of the four aggregation mechanisms. For each sketch, we asked participants to rate their confidence in the depicted weather forecast on a seven-point Likert item ("I feel confident that this will be tomorrow's weather.") ranging from "completely disagree" to "completely agree". Participants were aware that the shown forecast did not veridically display the weather of the next day. Subsequently, we asked participants to rate their preference for each sketch on three seven-point Likert items, each for using the depicted interface in one of the three possible scenarios (e.g., "I would like to use this representation for the scenario: Daily Dress Choice").

Participants

71 participants (37 male, 33 female) between an age of 19 and 77 years ($M = 24.8$, $SD = 7.0$) fully completed our online survey. We recruited the participants via mailing lists and social media channels of the University. The majority of our participants had completed a university degree (59.5 %) or a high school degree (26.8 %). At the end of the survey, participants had the possibility to enter a raffle of two 20 € Amazon vouchers.

70.4 % of our participants consulted weather forecasts multiple times a week. Their usage behavior changed according to season ("Especially in summer because the weather changes very often."), events ("I always check it before going to swim outside, and a barbecue, or a festival (everything outside where it could be fatal to wear the wrong clothes or the event is not appropriate for rainy weather)."), specific weather conditions ("Especially when the weather doesn't seem to be stable"), travel plans, and location.

53.1 % of our participants used more than two weather providers regularly ($M = 1.68$, $SD = 0.92$). Participants selected weather providers based on the perceived reliability and accuracy, usability, fast and easy access, and amount of information. Participants also stopped using weather providers because they either found the forecast inaccurate or the forecast was not easily available. Several reasons for comparing different sources were mentioned: to find out whether the sources provided consistent information ("If I really need to know the weather, I look at all of them to see if they match."), to compute the average ("I find they never agree with one another so I always try to find multiple sources and average them out (informally)."), or estimate the worst possible scenario ("I usually estimate the worst possible weather for a given scenario and prepare accordingly.").



(a) Confidence ratings for the aggregation mechanisms as a function of representation

(b) Preference ratings for the aggregation mechanisms as a function of representation

(c) Preference ratings for the aggregation mechanisms as a function of scenario

Figure 8.6: Participants' mean preference and confidence ratings (measured on a seven-point Likert-type item with 1 corresponding to completely disagree and 7 corresponding to completely agree); error bars represent standard errors of the mean. Figure 8.6c depicts centered mean ratings, i.e., the main effect of scenario has been removed to improve the perceptibility of the interaction effect.

Results

We performed two linear mixed-effects models analyses on the aligned-rank transformed Likert item data [Wobbrock et al., 2011] for participants' preference ratings and participants' confidence ratings. For the preference ratings, we used the fixed factors aggregation method, representation, and scenario as well as the random factor participant ID. For the confidence ratings, we used the same factors except for the factor scenario as we did not record participants' confidence on a scenario level. For significant effects, we conducted post hoc pair-wise comparisons with Bonferroni corrections.

Aggregation methods. Figure 8.6a shows participants' mean *confidence ratings* for each of the four aggregation methods as a function of the representations. Mean ratings were highest for *range* ($M = 5.34, SD = 1.46$), followed by *direct* ($M = 4.90, SD = 1.75$), *mean* ($M = 4.83, SD = 1.6$), and *single* ($M = 4.79, SD = 1.5$). The mixed-effects analysis revealed a main effect of aggregation method on participants' confidence ratings ($F(3, 1050) = 13.5, p < .001$). In particular, pair-wise comparisons showed that participants indicated higher confidence in *range* compared to the other three methods (range vs. direct: $t(1050) = 3.48, p < .01$; range vs. mean: $t(1050) = 4.85, p < .001$; range vs. single: $t(1050) = 6.0, p < .001$).

Range received the highest mean preference rating ($M = 4.85, SD = 1.73$), followed by *direct* ($M = 4.51, SD = 1.9$), *mean* ($M = 4.22, SD = 1.82$), and *single* ($M = 4.13, SD = 1.72$). Figure 8.6b shows participants' overall *preference ratings* for each of the four aggregation methods as a function of the representation. The mixed-effects analysis revealed a main effect of the aggregation method on participants' preference ratings ($F(3, 3290) = 37.78, p < .001$). Pair-wise comparisons showed that participants preferred *range* over the other three methods (range vs. direct: $t(3290) = 5.55, p < .001$; range vs. mean: $t(3290) = 7.69, p < .001$; range vs. single: $t(3290) = 10.21, p < .001$), as well as *direct* over *single* (direct vs. single: $t(3290) = 4.66, p < .001$).

These results support our hypothesis for RQ5.1 that people place higher confidence in aggregated data, especially if a *range* aggregation is used. Additionally, the results support our hypothesis for RQ5.2 that people generally prefer aggregated forecasts to single source forecasts.

Representations. *Line* received the highest mean preference rating ($M = 4.73, SD = 1.67$), followed by *pictogram* ($M = 4.70, SD = 1.73$), *bar* ($M = 4.42, SD = 1.81$), and *text* ($M = 4.28, SD = 1.88$). The mixed-effects analysis revealed a main effect of representation on participants' preference rankings ($F(3, 3290) = 16.48, p < .001$). Pair-wise comparisons indicated that participants preferred *line* and *pictogram* over *bar* and *text* (line vs. bar: $t(3290) = 4.17, p < .001$; line vs. text: $t(3290) = 5.92, p < .001$; pictogram vs. bar: $t(3290) = 3.682, p < .001$; pictogram vs. text: $t(3290) = 5.436, p < .001$).

Mean **confidence ratings** were highest for *pictogram* ($M = 5.17, SD = 1.53$), followed by *line* ($M = 5.09, SD = 1.49$), *text* ($M = 4.83, SD = 1.67$), and *bar* ($M = 4.70, SD = 1.69$). The linear mixed effects analysis revealed a main effect of representation ($F(3, 1050) = 7.27, p < .001$). Pair-wise comparisons showed that *pictogram* received higher confidence ratings than *bar* ($t(1050) = 4.08, p < .001$) and *text* ($t(1050) = 3.8, p < .001$).

These results indicate that participants preferred the representations *line* and *pictogram.* The latter also receives a confidence bonus.

Interactions. Of the three possible two-way interactions and the one three-way interaction between the three factors in the three-way analysis of **preference ratings**, only the interaction between *aggregation method* and *scenario* is significant ($F(6,3290) = 11.71, p < .001$). Figure 8.6c shows the mean preferences for each of the four aggregation methods as a function of the scenario. For illustration purposes, the data of each scenario was centered on the scenario's overall mean.

As the figure illustrates, a main source of the interaction effect is the increase in the preference ratings for the *direct* aggregation method as the importance of the scenario increases. Five of nine pairwise comparisons involving the *direct* aggregation method show significant differences of the difference of participants' preference ratings between the *direct* aggregation method and one of the other methods across two levels of the factor scenario (e.g., for the daily scenario *single* presentation is preferred over the *direct* aggregation while the situation is reversed for the wedding scenario; see Table 8.3 for test statistics and p-values for all significant contrasts). This partially confirms our hypothesis for RQ5.3 that the willingness to perform aggregation manually increases with the importance of the scenario (though *range* aggregation remains the overall preferred method).

A secondary source of interaction results from the stronger decrease in preference of the *single* source as compared to the other aggregation methods as importance of scenario increases (e.g. the difference between *single* and *range* is larger for the wedding scenario than for the daily scenario).

We find no interaction effect between aggregation methods and representations. In other words, there is not sufficient evidence in favor of our hypothesis for RQ5.4 that users have different preferences regarding the aggregation method depending on which representation is provided. Rather, users show an overall preference for *range*.

Qualitative Feedback. We additionally collected some qualitative feedback about the aggregation mechanisms. People mostly stated that the single source (especially the pictogram representation) "*seems like enough information for daily dress choice, and it is a similar representation to what [they] usually use. However, [they] would like a more detailed forecast for event planning.*"

As expected, participants preferred the direct comparison as they felt able to make an informed decision: "*I like having control and letting me make the decisions. I am informed of all possibilities, and it is up to me to come to a decision.*"

Table 8.3: Difference of the differences between two aggregation methods across two scenarios (e.g., is the difference between *direct* and *single* larger in the BBQ or in the Daily scenario?).

Contrast		**df**	χ^2	**p-value**
Direct-Single	**BBQ-Daily**	1	22.47	$< .001$
Range-Single	**BBQ-Daily**	1	10.35	.023
Direct-Single	**BBQ-Wedding**	1	11.37	.013
Direct-Mean	**Daily-Wedding**	1	25.24	$< .001$
Direct-Range	**Daily-Wedding**	1	10.23	.025
Direct-Single	**Daily-Wedding**	1	65.81	$< .001$
Mean-Single	**Daily-Wedding**	1	9.54	.036
Range-Single	**Daily-Wedding**	1	24.15	$< .001$

Participants also positively expressed that they liked the range aggregation: "*I like this representation because it provides almost as much information as the side-by-side comparison, but is much easier to see and interpret. It seems more trustworthy because it provides a range rather than a single value for temperature and rainfall. This acts as a margin of error, which means it is more likely to be correct.*"

For the mean aggregation, most participants stated that it "*does not represent the variation in forecasts between different weather providers, and it also (probably) does not provide the exact values reported by any one provider [and it] seems untrustworthy.*" In contrast to this, one participants also stated that he did not "*see the different temperatures and so on, but [he] know[s] that it's an average of some forecast sources, that's why [he] trust[s] the forecast.*"

Discussion

Participants generally preferred aggregation and their perceived confidence increased as well. However, the mean aggregation did not receive significantly better ratings than a single source forecast. Participants rather preferred range aggregation and direct comparison, which leave more room for an informed decision and provide more transparency on the actual value of the sources. The sheer knowledge of having multiple sources without knowing their actual values does not help to increase confidence.

Regarding the representations, participants preferred the line and pictogram representation. We assume that they were preferred due to users' familiarity with these representations as these are commonly used representations for weather data.

We did not find any interaction effect between aggregation mechanism and representation. This indicates that users do not prefer different aggregation mechanisms for different representations, but instead always prefer the range aggregation. Regarding our second evaluation, this allowed us to choose the aggregation mechanisms and the representation according to the overall preferences.

The interaction effect between the aggregation mechanism and the scenario indicates that with increasing importance of the scenario, users show a slightly increased preference for range and an increased preference for direct comparison, while the preference for single source forecasts drops below all other mechanisms. In important scenarios, people seem to prefer to have more control over their decisions by having more information and more work to aggregate multiple sources. Interestingly, this also applies for the range aggregation which seems to be a good compromise providing enough details to foster a sense of control.

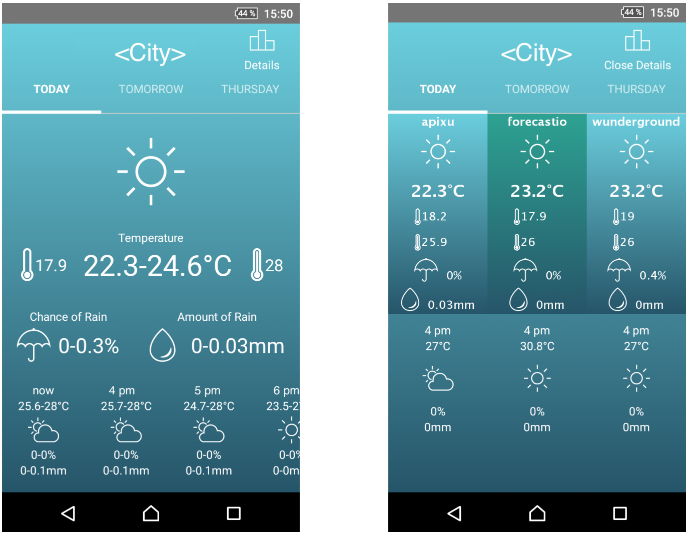
8.2.4 Evaluation in the Wild

Based on the previous results, we developed an Android weather application to evaluate the most promising aggregation mechanisms in the wild.

Method

We developed the weather application "Weather Compare" (see Figure 8.7). Based on the results of the online survey, we decided to show a pictogram representation and support range aggregation (see Figure 8.7a) and direct comparison (see Figure 8.7b). The application allows a user to toggle between the two aggregation mechanisms on demand, however, range aggregation is the default view. We picked three weather APIs that allow a reasonable number of API calls and are free to use for non-commercial/research purposes. Our application required users to turn on their location settings as the application automatically fetched the location of the user to show the forecast. We based the design of the application on the sketches shown in the user study, but added hourly information to make the application more informative.

All participants received detailed instructions on how the application works and an apk file to install the application on their personal mobile phones. We asked them to use the application for one week instead of their normal weather application.



(a) App - Range aggregation

(b) App - Direct comparison

Figure 8.7: Screenshots of the weather application "Weather Compare" with the main screen showing the range aggregation and the details screen showing the direct comparison of in total 3 different weather providers.

We logged participants' usage of the application and collected qualitative feedback after the course of the week on the (1) Future usage of the application, (2) Opinion about having data from multiple sources, (3) Usage of range vs. details view, and (4) Influence of aggregation on confidence.

Participants

We recruited 23 participants (12 male, 10 female, 1 preferred not to say) with an average age of 30.0 ($SD = 11.3$), who regularly use weather applications. We invited them by sharing the news about the application via social media channels of the university.

Results

During the week of using the application, participants opened the application on average 18.0 times ($SD = 5.7$). The details view was used in approximately one-third of cases ($M = 6.1, SD = 2.7$). In the following, we present the qualitative feedback of participants.

Future usage of "Weather Compare". All except three participants stated that they would like to use "Weather Compare" in the future. They, however, had specific feature requests such as including a wind forecast or a longer forecast range.

Opinion about having data from multiple sources. 14 participants explicitly stated that having data from multiple sources "*is the main thing that [they] liked about the app*". The aggregation helped them to better judge the forecast: "*I liked it because since this is a forecast, one provider is not really reliable but when you see that 3 different providers agree on something, it is more convincing.*" Only one participant did not appreciate the aggregation and stated "*that the cognitive load is too high when making the effort and actually comparing the values in contrast to just looking at a yellow sun or grey cloud.*"

Usage of range aggregation vs. direct comparison. Six participants stated that they "*always used the range representation and used the detailed view only out of curiosity.*" In contrast, three participants only used the details view. As one participant explained: "*I usually used the details view every day, because normally I would check at least 2 different sources to be sure about the weather, thus, I would like to know what each source says rather than seeing an average of them.*" One participant regularly used both views. Six other participants stated that if "*the range of values [was] too wide, [they] checked the details page.*" or "*before [they] spent a longer time outside*".

Influence of aggregation on confidence. Twelve participants stated that the application increased their confidence as they felt that the forecast was more detailed and having multiple sources made them trust it more. Only one participant stated that "*the range was typically only 2-3 degrees, in which case it does not influence [his] decision in any way.*" One participant did not think that having forecasts from multiple sources increased her confidence, "*because when they differ [she] kind of feel[s] like [she] lose[s] trust in all of them.*"

Discussion

With the help of the qualitative feedback, we were able to identify some shortcomings of "Weather Compare" arising from the fact that the APIs provided only a limited amount of information that we were able to access and show in the application. Participants missed certain features (e.g. typing in a location) or certain information (e.g. wind forecast) familiar from other applications. We assume that these shortcomings could be easily overcome by allowing users to individualize the application and offer more settings.

Regarding the APIs that we used, most participants had no strong opinion on what sources should be shown in the application. However, one participant explicitly stated that he wanted to choose the sources in the application. Providing more information about the used APIs and even a choice of providers for users could help them to understand constraints and increase their trust in the application.

Surprisingly, nine participants exclusively used one view instead of switching between views depending on importance, which we had expected to happen according to the results of the online survey. We assume that the weather forecast in general was of different importance to participants and that some did not want to give control to the application. In particular, participants who already compared forecasts before using *Weather Compare* commented highly positively about it.

In line with these findings, most participants stated that the aggregation of forecasts increased their confidence in the forecast. Interestingly, one participant felt that the application decreased her trust in weather forecasts in general. It seems that although research suggests otherwise, people might not be aware of the uncertainty in weather forecasts or perceive it lower than the actual uncertainty. It would be interesting to investigate whether this opinion changes after using the application long-term or when switching back to the weather application used before the study (as this is now maybe also perceived to be more uncertain).

8.2.5 Implications

Based on the results of our online survey and the in-the-wild evaluation, we derived five design implications for supporting the easy comparison of uncertain data. The implications reach beyond weather data and can serve as a starting point to develop applications supporting comparison of uncertain data in other application areas.

Support contexts with different importance: Applications should support range aggregation and direct comparison to allow the adaption to contexts with different importance. Although participants prefer aggregated data, they still want to make an informed decision and own judgements in very important scenarios.

Support perception at a glance: In everyday usage, users do not want to spend too much time in processing information and need a method to quickly make a decision by perceiving important information at a glance.

Support different types of users: Applications supporting multiple aggregation mechanisms should make switching between those easy and give people a choice to set their preferred default option. The evaluation in the wild showed that people have different tolerances for giving the control of aggregation to an application.

Give opportunities for choice: Applications should offer users a list of many providers to choose from, instead of pre-selecting the providers. This furthermore increases the feeling of control and potentially also trust.

Establish transparency: During the in-the-wild evaluation, participants requested additional forecast data (e.g. a wind forecast). By establishing transparency in providing detailed information on what information the providers actually offer, users would better understand the requirements and constraints of the application.

8.3 Humans' Internal Models for Aggregating Conflicting Uncertain Measurements

Wearable devices such as smartphones, smart watches, and activity trackers automatically contain a growing number of sensors. These sensors such as the accelerometer are often very small low-cost sensors to fit in the limited physical space. Additionally, specialized devices exist that are tailored to only track one specific property. All measured sensor data is uncertain as there might be measurement errors, the sensors might be wrongly calibrated, or the algorithms use thresholds to determine sensor values. As the number of personal general and specialized devices increases, the amount of conflicting and confusing information likewise increases.

The main goal of our work is to understand how humans aggregate conflicting probabilistic data. There are several mathematical ways of aggregating probabilistic data such as a simple average, a maximum likelihood estimator, or a winner-takes-all model [Ernst and Banks, 2002]. Budescu [2006] found that humans aggregate conflicting information from different sources by building an average. Their confidence in the aggregation depends on structural and natural

Table 8.4: Detailed research questions about how people interpret conflicting probabilistic information.

RQ	Research Question
RQ5.6	Does uncertainty information in general change how humans aggregate data?
RQ5.7	Do people make statistically more optimal decisions with different visualizations?
RQ5.8	Do people take the reliability of sensors into account?
RQ5.9	Do people make different decisions depending on different distances between measurements?

factors such as the amount of information, the number of experts judging the information, the inaccuracy, and the overlap of the information. However, it is unclear how the amount of presented uncertainty information influences humans' reasoning and their internal models. In the following, we present our detailed research questions and a user study investigating how humans aggregate conflicting sensor data.

8.3.1 Detailed Research Questions and Hypotheses

Our main goal is to understand how people interpret conflicting probabilistic information. We broke this goal down into five detailed research questions (see Table 8.4 for RQ5.6 to RQ5.9).

For RQ5.6, we assume that humans select the average if they have two point estimates, but use a different internal model if uncertainty information is provided. For RQ5.7, we assume that people do not make statistically optimal decisions. We, however, expect that people make statistically more optimal decisions if they have more uncertainty information. We further assume that a representation showing detailed aggregated uncertainty information as outlined in Section 7.3.1 (especially see Table 7.6), will lead to the statistically most optimal decisions. We assume that RQ5.8 is true and that humans adjust their aggregation towards the more reliable value, but will select the average if both values are equally likely. For RQ5.9, we assume that humans might change their strategy if sensor measurements differ too much.

8.3.2 User Study

In a user study, we evaluated four different representations combined with different weightings and different distances between estimates. We calculated the weightings based on a formula for calculating the weighted average of two measurements [Taylor, 1997]:

$$X_{est.} = \frac{w_A\, Measurement_A + w_B\, Measurement_B}{w_A + w_B} \tag{8.1}$$

where the weight of each measurement is determined by the normalized reciprocal variance of the measurement:

$$w_A = \frac{1}{Var_A} \quad and \quad w_B = \frac{1}{Var_B} \tag{8.2}$$

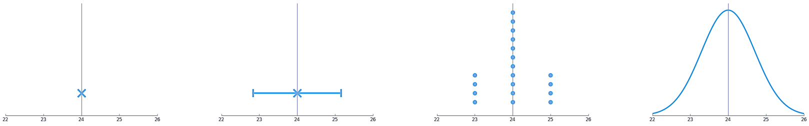
Design

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 distribution function), *weighting* (with the five levels 50 %-50 %, 40 %-60 %, 30 %-70 %, 20 %-80 %, 10 %-90 %), and *distance* (with the three levels no distance, small distance, large distance). Besides the details about the current condition, we only logged participants' answers.

Visualizations. We used four different visualizations based on the amount of uncertainty included in a representation defined in Section 7.3.1. We only used graphical representations to increase their comparability. We used a point estimate without uncertainty information, a confidence interval with aggregated uncertainty information, a dotplot (as suggested by Kay et al. [2016]) for detailed aggregated uncertainty and a probability distribution function for detailed uncertainty.

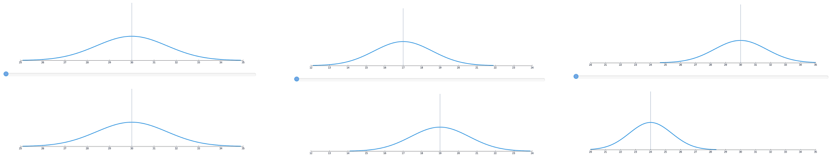
Weighting. We chose the associated variances for the measurements such that one weight increased in steps of 0.1 from 0.5 to 0.9 and the other weight decreased from 0.5 to 0.1 (for both weights after normalizing). In the following, we refer to these weights as 50 %-50 %, 40 %-60 %, 30 %-70 %, 20 %-80 %, and 10 %-90 %.

Distance. We used zero distance and two different distances between the shown measurements: a small and a large distance (see Figure 8.9). The small and large distances equaled on average 1.9 and 5 times the size of the standard deviation of the distribution used in the 50 %-50 % weighting condition.



(a) Point estimate (b) Confidence interval (c) Dotplot (d) Probability distribution function

Figure 8.8: The four visualizations used in the user study each including a different amount of uncertainty information as defined in Section 7.3.1.



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

Figure 8.9: We used three different distances between measurements in our experiment. This figure shows an example for each distance.

Method

At the beginning of the study, participants had to fill a demographic questionnaire before the study instructor explained the general task. For each of the four visualizations, participants were asked to select the true value on a slider that was shown in the middle of two visualizations showing different values. We used a Latin square design to determine the order of the visualizations for each participant. Before answering the first task with a new visualization, participants got a short explanation of the visualization and a trial period to get acclimated to the visualization and the interface. We additionally randomized the placement of the visualizations on the page (top-bottom and left-right). Participants experienced each weighting 40 times per visualization, which resulted in a total amount of 200 trials per visualization and 800 trials overall. At the end of the study, participants filled a Berlin Numeracy Test [Cokely et al., 2012] to assess their statistical knowledge.

Participants

16 participants (8 male, 8 female), with an average age of 24.6 ($SD = 3.4$) participated in the study. Most of the participants were students as participants were recruited in a university setting. The Berlin Numeracy Score of participants ranged from 3 to 7, showing a basic but not necessarily an expert understanding of statistics.

Results

Based on participants' answers, we calculated the weights that they assigned to the two presented measurements and calculated the difference to the actual weight that we used to generate the visualizations. For each trial, we used the weight difference of the larger weighting for the analysis. For example, if a participant saw two measurements with a weighting of $20\,\%$-$80\,\%$ and instead internally used a weighting of $30\,\%$-$70\,\%$, the weight difference would be $10\,\%$ or 0.1 for the larger weight. We performed a linear mixed-effects model analysis on the aligned-rank transformed data [Wobbrock et al., 2011] for the weighting differences as the data was not normally-distributed. We used the fixed factors visualization, weighting, and distance as well as a random factor for the participant ID. For significant effects, we conducted post hoc pair-wise comparisons with Bonferroni corrections.

The mixed effect analysis revealed a main effect of the visualization, the weighting, and the distance on the weighting difference. However, all two-way interactions and the three-way interactions are as well significant which might impact the main effects (see Table 8.5 for details).

Visualization. The post hoc test revealed that the weight differences of all visualizations differ significantly except the difference between the dotplot and the probability distribution function. The average weight difference was highest for the point estimate ($M = 0.17$, $SD = 0.18$), followed by the confidence interval ($M = 0.06$, $SD = 0.24$), the probability distribution function ($M = 0.05$, $SD = 0.26$), and the dotplot ($M = 0.04$, $SD = 0.24$). These results support the hypothesis that uncertainty information changes how people in general choose the true value. We can additionally partially support our hypothesis that people make statistically more optimal aggregations with uncertainty information, however, the dotplot was not significantly better than the probability distribution function.

Weighting. The post hoc test revealed that the weight differences between all weightings differed significantly. The weight difference was smallest for

Table 8.5: Results of the linear mixed-effects model analysis on the aligned-rank transformed data [Wobbrock et al., 2011] for the three main factors visualization, weighting, and distance, and the random factor participant ID.

Effect/Interaction	df	df.res	F	p-value
Visualization	3	12113	1201.969	$< .001$
Weighting	4	12113	692.652	$< .001$
Distance	2	12113	787.455	$< .001$
Visualization:Weighting	12	12113	170.183	$< .001$
Visualization:Distance	6	12113	215.389	$< .001$
Weighting:Distance	8	12113	175.496	$< .001$
Visualization:Weighting:Distance	24	12113	54.775	$< .001$

the 50 %-50 % weighting ($M = -0.01$, $SD = 0.19$), followed by the 40 %-60 % weighting ($M = 0.04$, $SD = 0.22$), followed by the 30 %-70 % weighting ($M = 0.07$, $SD = 0.23$), followed by the 20 %-80 % weighting ($M = 0.12$, $SD = 0.25$), with the highest weight difference for the 10 %-90 % weighting ($M = 0.18$, $SD = 0.26$). The higher the weighting, the bigger the difference of participants to statistically optimal judgements. These results support our hypothesis that people selected the average for the 50 %-50 % weighting and adjusted the value towards the more reliable value for different weightings.

Distance. The post hoc test revealed that the weight differences between all distances differ significantly. The weight difference was smallest for the zero distance condition($M = 0.00$, $SD = 0.00$), followed by the large distance ($M = 0.09$, $SD = 0.26$) condition. The weight difference was highest for the small distance ($M = 0.10$, $SD = 0.25$) condition. These findings support our hypothesis that people select the average for no distance between values, however, the distance has an unexpected influence on the selection of the true value.

Interaction between Visualization and Weighting. Figure 8.10a shows the mean weight differences for each of the four visualizations as a function of the weighting. The post hoc test revealed that the differences of differences between the point estimate and all other visualizations are significant. For the less equal weightings (20 %-80 %, 10 %-90 %), the difference between the dotplot and the probability distribution function to the confidence interval is significant. This does not contradict the main effects.

Interaction between Visualization and Distance. Figure 8.10b shows the mean weight differences for each of the three distances as a function of the visualization.

All differences are significantly different expect from the difference between the zero distance and the large distance condition for the dotplot and the probability distribution function. However, the difference between the dotplot and the probability distribution function differs significantly between small and large distance. This does not contradict the main effects, but provides an explanation for the difference between small and large distance.

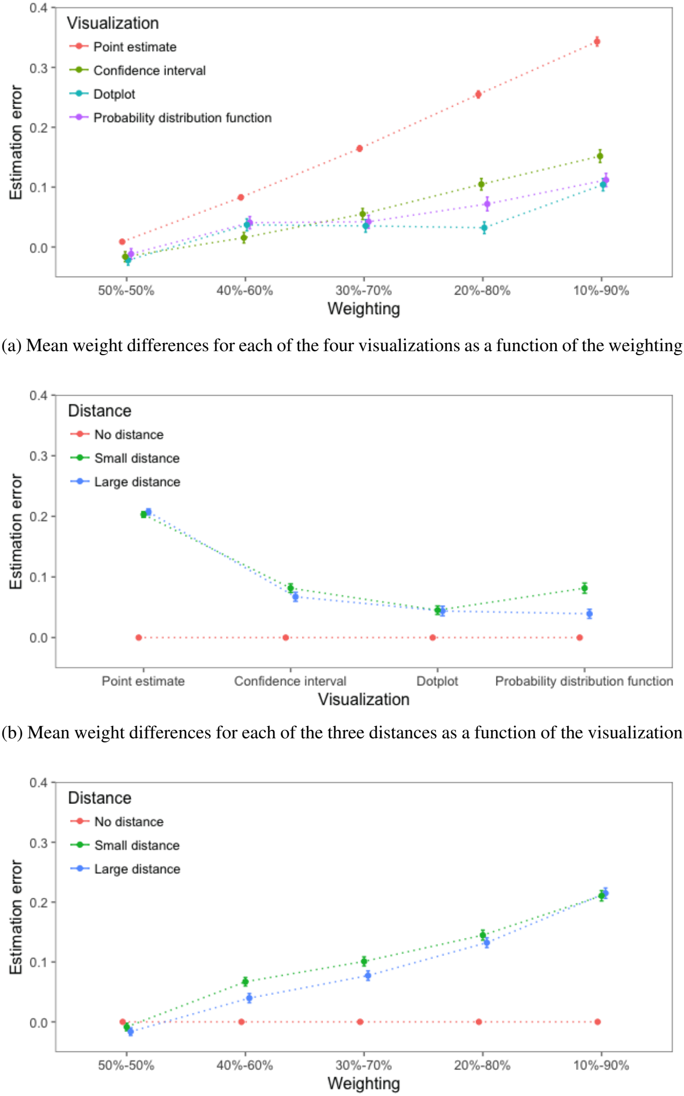
Interaction between Weighting and Distance. Figure 8.10c shows the mean weight differences for each of the three distances as a function of the weighting. The post hoc test revealed that the differences of differences between the zero distance condition and the small or large distance conditions are all significant. For the most equal and most extreme weightings (50 %-50 %, 10 %-90 %), there is no significant difference between the small and the large distance condition. This also does not contradict the main effects, but provides further insights in the difference between small and large distance.

Three-way Interaction. Figure 8.11 shows the three distances as a function of the weighting separately for each of the four visualizations. These diagrams most clearly underline the following finding:

1. With information about sensor reliability (see Figure 8.11a), participants chose the average value,
2. The higher the difference in the sensor reliabilities, the higher participant's estimation error,
3. Uncertainty information lowers the increase of the error,
4. For larger differences in sensor reliabilities, dotplot and probability distribution are most suited, and
5. The error is larger for small inconsistencies, in particular for the probability distribution function.

8.3.3 Discussion & Implications

Our results show that displaying uncertainty information changes humans' internal models of how they aggregate two distinct values. As expected, humans choose the average if they have no uncertainty information, but weight the values as soon as they have uncertainty information. More detailed information in form of a dotplot or a probability function makes the weighting more optimal than showing a confidence interval.

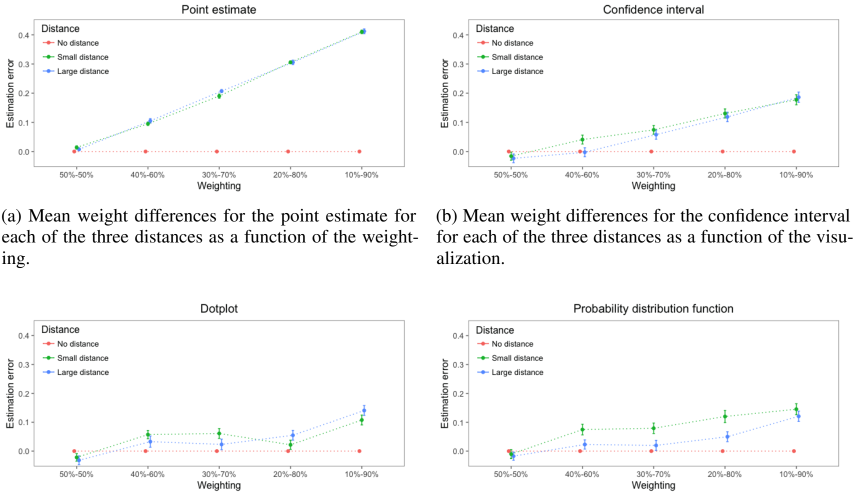


(a) Mean weight differences for each of the four visualizations as a function of the weighting.

(b) Mean weight differences for each of the three distances as a function of the visualization.

(c) Mean weight difference for each of the three distances as a function of the weighting.

Figure 8.10: Participants' mean weight differences for combinations of two variables; error bars represent the standard errors of the mean.



(a) Mean weight differences for the point estimate for each of the three distances as a function of the weighting.

(b) Mean weight differences for the confidence interval for each of the three distances as a function of the visualization.

(c) Mean weight difference for the dotplot for each of the three distances as a function of the weighting.

(d) Mean weight difference for the probability distribution function for each of the three distances as a function of the weighting.

Figure 8.11: Participants' mean weight differences plotted for each visualization; error bars represent the standard errors of the mean.

For the different weightings, our results show that the higher the deviation between the two weightings, the bigger the error to the statistically optimal judgement. People are more conservative and do not choose extreme weightings such as 10 %-90 %. In combination with the visualizations, we can conclude that for nearly the same reliability, showing uncertainty information makes no sense as there is no advantage of showing this information and point estimates work well to calculate an average. The larger the difference between the reliability of sensors, the more participants benefitted from additional uncertainty information to make statistically more correct judgements. The error of the weighting grew linearly with the point estimate and also for the confidence interval, we can see a linear trend. The confidence interval, however, might still be applicable for cases with a weighting between 40 %-60 % and 30 %-70 %.

For the distance, we found that for zero distance between point estimates, participants always selected the average no matter what visualization or weighting was used. Here, a point estimate can be enough to communicate the information as it will produce the same result. For the small and the large distance condition, we, however, experienced that participants made larger errors with the smaller distance. We assume that for smaller distances, people were biased towards the average while this bias was not in place for larger distances. It is also interesting that the probability distribution function seems to produce larger errors in comparison to the dotplot for the smaller bias. In Figure 8.11c it becomes clear why there is this difference. As the dotplot performed slightly better for some weightings with the small distance and for others with the large distance, the effect cancels out for the two-way interaction. However, the values indicate that the difference between the small distance and the large distance rises with the additional uncertainty information. This might be a slight indicator that the dotplot should be preferred to a probability distribution function for communicating uncertain information.

8.4 Insights for Supporting the Interpretation

In this chapter, we took a closer look at the interpretation of uncertain data through the user. We mainly explored humans' internal models when making predictions, and when aggregating conflicting information.

We found that user-made predictions are a suitable tool to support humans in understanding their behavior patterns and learn how such predictions can be improved. We argue that these insights could be an interesting start to use user-made predictions to make complex algorithms predicting user behavior easier to understand as they may use the same principles to make their predictions.

For the interpretation of conflicting uncertain data, we identified five design implications to design for comparison. The main findings are that differences in users' character and their judgement of scenarios play an important role as identified in Chapter 5. Additionally, transparency of information sources needs to be established to foster trust.

We further found that uncertainty information improves users' aggregation of conflicting information if uncertain data from multiple sources is presented. Participants use a weighted average although they do not make a statistically optimal aggregation. The larger the difference between the reliability of uncertain data, the more uncertainty information helped participants to make better aggregations. For non-conflicting information, point estimates are a suitable way of visualization the data.

Based on all our previous findings, we present our conclusions in the next part of this thesis. We created a web-based simulation tool for end-users that allows researchers to experiment with the input and output methods developed in the context of this thesis. The use case of end-user simulations is used as a sample use case for uncertainty communication in interactive systems.

IV

Conclusion and Future Work

Chapter 9

Simulation Tool for End-Users

We implemented a web-based simulation tool called SimulaTE including input methods presented in Section 6.1 and Section 6.2 as well as output methods presented in Section 7.3. The tool incorporates these input and output methods for uncertainty to allow researchers and the general public to explore different input and output methods for calculations and small simulations. In this chapter, we summarize the functionality of the simulation tool and the user-centered process we followed during its development.

9.1 Implementation

The simulation tool was built with Django 1.8.6[8]; a python-based web framework and Bootstrap[9] as a front-end framework. The tool is mainly divided into two distinct parts: the toolbox to build modules and models, and the simulation part to run models. Users need to sign up or log in to be able to access the toolbox and the simulations.

An overview of the general architecture of SimulaTE is provided in Figure 9.1. The module creator allows users to create their own modules, which are stored in

[8] Django: `https://www.djangoproject.com/start/overview/`

[9] Bootstrap: `http://getbootstrap.com/`

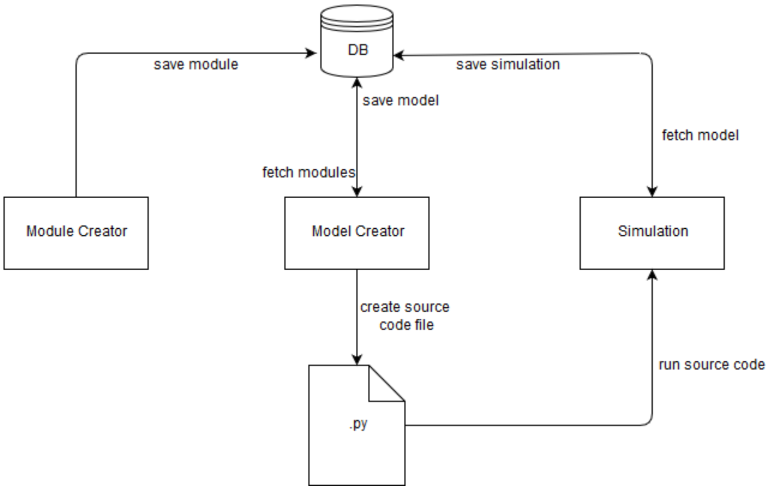


Figure 9.1: Overview of the three main components of SimulaTE and their connections.

a database and used to create models. The model creator allows users to create models. As models consist of modules, the model creator fetches the existing modules from the database. For each model, a Python file is created and some additional information is stored in the database. The simulation component runs the Python file for given inputs and fetches some additional model information from the database. The results of all simulation runs are also stored in the database.

9.2 User-Centered Process

To develop the end-user simulation tool, we used a user-centered process. We incorporated multiple simulation experts from the Cluster of Excellence in Simulation Technology in our design process by conducting interviews with them in different steps of the development process. All of the experts were from the Social Sciences and without any programming knowledge. We incorporated their feedback directly into our implementation by adapting the interface to make it easier for non-programmers to use.

We further conducted a small usability study with the first functional prototype of the tool. We ran a pilot study with one participant to make sure that our task descriptions were understandable and adapted them based on the feedback. We then recruited four participants (3 male, 1 female). On a scale from 1 to 5 where 5 corresponds to an expert programmer, participants reported their programming experience to be at the lower end ($M = 2.0, SD = 0.82$). Participants had to solve tasks by using the simulation tool. The tasks included sign up, creating an easy module, creating a more complex model for an existing module, and running simulations. We did screen and audio recording and asked the participants to think aloud. One study instructor documented their actions during the study to identify potential usability flaws. The participants mainly struggled with the parameters of the inputs and outputs, so we adapted the respective parts based on their feedback to improve the overall usability of the simulation tool.

We implemented several modules and models as examples: First, we implemented some mathematical formulas such as the quadratic formula. Second, we implemented small models calculating the interest on a sum of money over time and the ecological footprint for different means of transportation. As a proof of concept, we additionally implemented an import for Fitbit[10] data and an import for a food database to build models related to health and calorie intake.

9.3 Functionality

SimulaTE consists of a number of different pages. Besides the main page, which contains an explanation of the whole tool and of the process to create and run simulations on the platform (see Figure B.1), the tool contains an overview of all available input methods, all available output methods, a module creator, a model creator, and a simulation environment. Figure 9.2 shows a use case diagram of the tool, which includes an overview of all functions. In the following sections, we will explain these functions in more details.

9.3.1 User Accounts

The main page and the descriptions of input and output methods are visible for every visitor to the tool's webpage. However, the module creator, the model

[10] Fitbit: `https://www.fitbit.com/de/home`

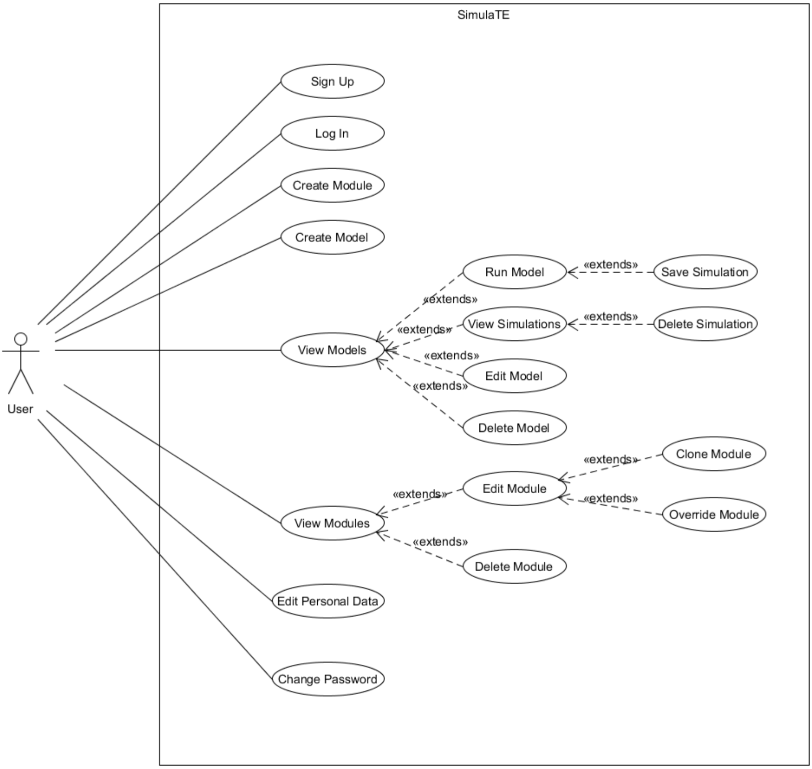


Figure 9.2: Use case diagram for SimulaTE.

creator, and the simulation environment are only accessible for registered users. Users can either sign up to create a new account for the tool or log in with an existing social media account. On the personal profile, users can edit personal information and change their password. Modules are personal building blocks that are not shared between users; however users can decide for models whether they should be private or public. Simulation runs are also stored on a personal level and not shared with other users.

9.3.2 Modules

Modules are small building blocks that have a name, a description, an abstract set of inputs and outputs, and a function. Figure B.2 shows the module creator which allows the user to create new modules from scratch. The inputs and outputs only have a name and a description, but are not yet coupled to specific input and output methods. The function of a module needs to be implemented in Python code. The function is separated from the input and output methods to allow the same function to be used in multiple models with different input and output methods, for example classical methods or methods that allow the users to enter or show uncertainty.

Figure B.3 shows the modules overview in a users' profile. Modules are private and always belong to one user account. A user can edit or delete modules in the personal overview. If a module is edited, the user can decide to either override the existing module or clone the module before editing it to avoid overriding modules already used in models.

9.3.3 Models

A model in SimulaTE consists of one or several modules. Models can be created in the model creator (see Figure B.4). The user can simply add or remove modules from the model by clicking on them, and can then choose specific input and output methods for the concrete model via drop down menus. The input and output methods can additionally be adapted to personal needs by specifying parameters, which are further explained in the input and output overview. A model additionally has a name, a description, an optional picture and can have user-defined tags. In contrast to modules, a model can be private or public. Logged in users can see and execute public models.

Figure B.5 shows the model overview. In this view, users can edit or delete their own models. Additionally, users can run their own or public models and access all saved simulation runs previously executed for a model. A user can toggle between the page showing all models, owned models, or all models that have already been used to run simulations. To search for models, the user can enter either free text or tags in the search bar on top of the page.

A click on the edit button of a model navigates the user to the model editor (see Figure B.6). The editor shows either an abstract editor or a code editor to allow more experienced users to adapt the program code. It is possible to switch between the two interfaces during the creation process of a model.

9.3.4 Input Methods

SimulaTE supports a whole range of input methods with different properties. The input methods page provides an overview of all input methods that are supported by SimulaTE. Figure B.7 shows a part of the input method page showing the sliders that are described in Section 6.2. For each input method, we provide an example where the users can try out the input method. Additionally, we list all parameters of the input method and explain them in more details. Parameters allow users to manipulate the input methods to match their needs, for example to show default values, preselect valid ranges, or define descriptions of input fields.

SimulaTE offers input methods from the following four categories:

- **Simple:** Contains input methods that allow to enter a single value, such as a number, a Boolean value, or text.
- **Percentage:** Contains all input methods that allow to enter a percentage value.
- **Range:** Contains all input methods that allow a user to enter a number range.
- **Others:** Contains all specific input methods that do not fit into the other categories, e.g. sensor connections that allow to automatically transfer data from sensors.

In all categories, SimulaTE offers both standard input methods and input methods that allow users to explicitly enter uncertainty. Models can support both ways of input to actually allow users to decide whether they rather want to give deterministic or uncertain input.

9.3.5 Output Methods

SimulaTE also supports a whole range of output methods. The output method page provides an overview on all output methods that are supported by SimulaTE, for example the visualization depicted in Section 7.3. Figure B.8 shows a part of the output page with an area chart. As for the input methods, the page contains one example of each output method and an overview of the parameters that can be used to adapt the methods to the user's personal needs. SimulaTE provides different output methods such as bar charts, histograms, area charts, pie charts, as well as simple text output.

9.3.6 Simulations

A simulation in SimulaTE is equivalent to the execution or run of a model. With a click on the run button, the user enters a page that can execute the model. Before execution, the user has to specify the necessary input parameters (see Figure B.9) to get to the presentation of the output. After running a simulation, the user can save the run to later have another look at the results.

The user can also navigate to all saved simulations for one specific model (see Figure B.10). For each saved run, the user can have a look at the input values and the output values.

9.4 Implications

We implemented a web-based simulation tool called SimulaTE, which supports end-users in building and running calculations and simulations. We used a user-centered process to implement and refine the tool incorporating simulation experts and end-users with little programming knowledge.

Our development was based on the requirements identified in Chapter 4. We built a general tool that allows users to run very distinct calculations and simulations. We separated the model building step from the simulation step as public models can be run by any user. Furthermore, we minimized necessary mathematics and programming knowledge wherever possible. As functional requirements, we included different sources of parameter input by allowing users to directly enter data, connect devices, or use databases. We implemented an example model using Fitbit data and a food database for a proof of concept. The tool further supports different illustrative visualizations in addition to the output of plain numbers. Our main focus when building the tool, however, was on the non-functional requirements of flexibility and transparency. To increase flexibility, modules can be part of multiple models using different input and output methods. To foster transparency, we especially offer input and output methods suitable for communicating uncertainty.

The tool offers a starting point for future research in the direction of uncertainty in interactive systems. Researchers, practitioners, and the general public can use the tool to explore input and output methods communicating uncertainty.

Chapter 10

Conclusion

In this thesis we have systematically explored the occurrence and handling of uncertainty in the context of interactive systems. We showed that uncertainty plays an important role in interactive systems and developed novel input and output methods that support users in effectively dealing with uncertainty.

10.1 Research Contributions

Based on related work, we identified sources of uncertainty in interactive systems by enhancing the General Interaction Framework. We then explored three important key aspects: the input, the output, and the interpretation of uncertainty. On the input side, we contribute methods for the explicit and implicit input of uncertainty. We explored the design space for standard input controls, specialized slider controls, and tangible input controls. We additionally showed the feasibility of using selected behavioral and physiological measurements to implicitly capture uncertainty. On the output side, we found a lack of communication of uncertainty in current mobile applications though users voiced a preference for it. We contribute designs to improve the communication of uncertainty for activity tracking data, and further contribute a classification for representations based on the amount of uncertainty information communicated. We show that representations showing detailed aggregated uncertainty information are most promising to use when communicating uncertainty to the general public. Regarding the interpretation, we found that engaging users to make predictions could improve their reasoning about uncertainty. Additionally, we contribute design recommendations for showing conflicting information in a weather context and for general measurements. In the first case, we focused on users' confidence and in the second case on internal models that humans use to aggregate information. We included the developed input and output methods in an end-user simulation tool to simplify future research and usage of these methods.

In the following, we provide details on the findings for the research questions identified at the beginning of this thesis.

10.1.1 Current Simulation Usage

In **RQ1**, we asked **What can we learn from the current usage of simulations?** We conducted multiple small research probes to find answers for this question.

From analyzing expert simulation usage, we learned that simulation tools are very specialized tools mainly used for one application area. Additionally, different tools are used in different steps of the simulation process which might even be carried out by different experts. To make simulations usable for the general public, a more general approach has to be followed as less mathematics and programming knowledge can be expected.

From non-expert usage, we learned that plenty of use cases exist for simulation usage in everyday life, e.g. related to health or finances. We identified a set of four main functional requirements for an end-user simulation tool: (1) support of different sources of parameter input, (2) illustrative visualizations, (3) support of short- and long-term predictive simulations, and (4) context-awareness. We additionally identified two non-functional requirements: flexibility and transparency. To reach flexibility and transparency, uncertainty has to be taken into account at all levels of the simulation process.

10.1.2 Sources of Uncertainty in Interactive Systems

In this section we focus on **RQ2: What are the sources of uncertainty in interactive systems?**

Based on related work, we identified 13 distinct sources of uncertainty in interactive systems, and integrated them into the General Interaction Framework.

Sources of uncertainty were identified for the user, the input, the system, the output, and all connections between them. Some sources, however, might apply to multiple components of the General Interaction Framework. Sources of uncertainty related to the system are not very relevant for the HCI community as they are part of specific sciences, and sources of uncertainty related to the user are part of research in Psychology; however sources of uncertainty connected to the articulation, input, output, and observation are of specific interest for the HCI community.

By developing novel input and output methods for interactive systems dealing with uncertain data, uncertainty can either be reduced or quantified to appropriately take it into account in all stages of system usage. New input techniques allowing one to enter uncertainty can help to overcome uncertainty hidden from a system due to a lack of user knowledge or imprecise measurements. The system can take these sources into account as soon as a user communicates this uncertainty to the system. Evaluations of new controls can help to minimize the sources of uncertainty related to limited understanding of the users. Novel input and output methods are also opportunities to provide more degrees of freedom and accuracy than current input controls to further reduce uncertainty. Evaluations of new output controls can help prevent future misjudgments.

We conclude that uncertainty plays an important role in interactive systems, although it has often been neglected in past research. More research in the area of uncertainty can help designers and researchers to better design for this aspect.

10.1.3 Input Methods

In the following we discuss **RQ3: What input controls are suitable for uncertain input?** To answer this we conducted several research probes to explore different design spaces related to input methods.

We considered multiple areas for the input of uncertain data: standard input controls, specialized slider controls, tangible input controls, and behavioral and physiological measurements. We showed that explicit and implicit methods are feasible and suitable for quantifying uncertainty in user input.

For standard input controls, we suggest using an additional number field or slider to allow users to specify a probability percentage. This is a small change compared to current interfaces which allows users to indicate their uncertainty when using a system. For a solution that adds further transparency and more degrees of freedom, we propose specialized slider controls; the probability distribution sliders. One advantage of these is that the slider controls allow for matching the degrees of freedom to the context, the task and the statistical knowledge of the user. The attached visualizations provide additional transparency on how the system interprets the input data. We further identified tangible shape-changing devices as a promising research area for input with uncertainty. The shape-changing aspect of an interface can be leveraged to offer a smooth change between input with and without uncertainty. This strengthens the connection between the single-value input and the additional uncertainty value. Based on our evaluations, we propose to use the SplitSlider design for the input of uncertain data.

For the implicit input of uncertainty, we recommend a combination of eye tracking and key logging. However, more advanced physiological measurements might in the future lead to better results.

10.1.4 Output Methods

We further contribute findings for **RQ4: What visualizations are suitable for uncertain output?** To answer this question we conducted multiple research probes to analyze the current state and compare different representations for uncertain data.

We first analyzed how current mobile applications communicate uncertain data and found that most applications do not communicate uncertainty at all. However, users voiced a preference for uncertainty communication.

We then created three designs for communicating uncertainty in activity tracking data. We propose to communicate a range instead of a fixed value and add grey bars or other visual elements to bar charts that indicate the uncertainty. Our studies showed that although users prefer to have more information, it still has to be quickly graspable. Thus, adding too much complexity by communicating uncertainty has to be avoided to minimize cognitive load.

We further found that visualizations showing more uncertainty information are not necessarily more suitable than visualizations with less uncertainty information. Familiarity, visual appeal, and the ease to understand a visualization are important influencing factors. We contribute a classification of visualization by the amount of uncertainty information that they include. The classification can support researchers in selecting representations to support decision-making. We found that representations with aggregated uncertainty information gave participants a misleading feeling of control. We therefore do not recommend to use such representations, but rather focus on visualizations that include detailed aggregated uncertainty information such as a histogram or dotplot. In our study, the histogram proved to be most suitable for supporting users in decision-making.

10.1.5 Interpretation

In this section we focus on our research question **RQ5: How do people interpret uncertain data?** We conducted three research probes with different main foci to better understand how people interpret uncertain data.

We found that user-made predictions can improve users' understanding of their personal behavior. Participants in our study started to discover behavior patterns and reason about their daily activities to be able to improve their predictions. They also had a strong will to improve their predictions and tried to reach their predictions if possible.

For the interpretation of conflicting data, we found that a computationally generated average will not increase confidence in uncertain data. Participants rather preferred a range view or details view as they wanted to make their own judgements. To design for comparison, different types of users and scenarios have to be taken into account. Additionally, transparency of the sources and the gathered information establishes trust. We recommend to support easy comparison with a range aggregation and a direct comparison to allow users to switch between the two modes based on the context and their preference.

We further showed that uncertainty visualization improves reasoning about conflicting data. Point estimates are suitable for presenting non-conflicting data or data with equal reliability as humans tend to select the average. If this is not the case, uncertainty information can help users select a weighted average closer to a statistically optimal value. We suggest a dotplot for supporting users in aggregating conflicting information.

10.2 Concluding Remarks

Humans are confronted with uncertainty whenever they make a decision and more uncertainty is also introduced in the digital world. Interactive systems include uncertain information originating from machine learning or predictive algorithms.

As past studies in Psychology revealed the positive effects of communicating uncertainty, such as increased trust and more optimal decision-making, it should be the aim of HCI researchers and designers to adequately communicate uncertainty to their users.

This thesis has provided a systematic exploration of uncertainty in interactive systems. Based on the identification of sources of uncertainty, we implemented and evaluated novel input and output methods for communicating uncertainty. We are convinced that this is an important step to bring uncertainty into people's minds and raise awareness for the importance of uncertainty communication in the HCI community. The interest in the workshop "Designing for Uncertainty in HCI", which we organized at CHI'17, shows that the topic has recently gained attention in the HCI research community, which was not the case when we started our exploration at the end of 2013.

We hope that our work and our developed tool can support and inspire practitioners and other researchers to include uncertainty quantification and communication in their design and development process. We outline potential follow-up experiments and future work in the last chapter of this thesis.

Chapter 11

Future Work

This thesis provides a systematic exploration of uncertainty in interactive systems to set a common ground and starting point for future research about uncertainty in HCI. During the course of this thesis, additional research questions appeared that are beyond the scope of this thesis. This chapter summarizes interesting follow-up research questions for future work.

11.1 Sources of Uncertainty

The presented classification of sources of uncertainty in interactive systems is a first step to raise researchers' and developers' awareness of these sources. As a next step, it would be helpful to further extend the classification to make it more concrete.

The classification could be extended by adding potential types of uncertainty associated with the sources and more concrete suggestions and guidelines for researchers and developers on how to handle uncertainty. This thesis mainly focused on suggesting new ways to increase the degrees of freedom in the input and output methods. We additionally offered suggestions on how to avoid misjudgments. Interdisciplinary work with psychologists, simulation experts, or domain experts that carry out uncertainty quantification could help to further understand sources of uncertainty beyond was is of primary relevance to HCI.

It would additionally be helpful to collect best practices and concrete application scenarios for interactive systems dealing with uncertainty. As the topic will increasingly gain more attention in the HCI community, concrete application scenarios with associated sources and explanations on how they handle uncertainty could help researchers and developers to better understand which sources apply for their concrete application scenario. Several methods such as literature reviews, interviews with researchers and developers, and developing concrete best practices together with these could be methods to apply for future research.

11.2 Input Methods

We explored different design spaces of input methods for entering uncertain data, especially standard input controls, sliders, tangible interfaces, and physiological measurements. However, other input methods remain to be explored. We suggest to further explore different input methods such as gesture input, voice input, or brain computer interfaces to understand whether users can be enabled to enter uncertainty with such interfaces.

As we mainly tested our functional prototypes in lab studies, it would be interesting to include the input methods in real-world systems. Specific research prototypes could actually collect and measure the data to compare how well users can estimate the uncertainty. Based on this data, models for uncertainty in user input could be derived.

We also only tested very specific application scenarios with our input methods. All experiments could be repeated within different application scenarios to understand whether the findings can be easily transferred to other scenarios.

11.3 Output Methods

For the output, this thesis mostly focused on the graphical representation of uncertainty. Further experiments could compare graphical representations to purely linguistic and numerical presentation for different application scenarios in interactive systems.

The same that applies to the input methods also applies to the output methods for communicating uncertainty; as we mainly conducted lab studies, it would be interesting to include the output methods in real-world systems to understand whether the findings transfer to different application areas. In addition to real-world systems, further games with different context could also help to better understand how much findings depend on the presented scenarios.

11.4 Interpretation

Our exploration on user-made predictions could easily be continued. This could include conducting follow-up experiments on the degree to which predictions influence humans' dealing with uncertainty and whether predictions could influence behavior change. Long-term studies would be needed to further explore this.

The psychological aspect of aggregating conflicting information could easily be extended. The work should be repeated with different contexts. Furthermore, it would be interesting to understand whether naming specific sensors and framing the question differently would influence users. To better understand the influence of biases, users could even be handed real sensors to carry for several weeks to collect and aggregate measurements. Further research could also construct mathematical models for how humans aggregate data based on different representation alternatives.

Another interesting research topic could be to understand how humans aggregate conflicting news or other textual information to offer support strategies other than numerical data. Humans might apply very different internal models for other data types.

Another question is on how much training influences the interpretation of statistical data. Statistics education could make a difference on how humans interpret uncertain data. Specific training methods might be suitable to remove biases and help humans to make more optimal decisions under uncertainty.

11.5 End-User Simulation

The simulation tool SimulaTE offers a starting point for researchers to use input and output methods for communicating uncertain data. The tool can be easily extended to support additional functionality. It could for example be extended to not only offer the possibility to create models and run simulations, but also allow researchers to analyze how users run simulations with public models. The tool could additionally be connected to more external data sources and data bases to support different kinds of simulations.

11.6 Concluding Remarks

With this thesis, we set a starting point for the research on uncertainty in interactive systems. Although we have identified many open questions for future research, the thesis still contributes a number of important findings to the research community. We identified uncertainty in interactive systems as an important research topic by applying insights gained from research of other fields to the HCI context, and used these in the identification of sources of uncertainty. We furthermore developed suitable input and output methods, which helped to identify guidelines and further ideas. We built a system that includes this input and output methods to support future research and generate further insights.

V

Bibliography

BIBLIOGRAPHY

Jeroen C. J. H. Aerts, Keith C. Clarke, and Alex D. Keuper. Testing popular visualization techniques for representing model uncertainty. *Cartography and Geographic Information Science*, 30(3):249–261, 2003. doi: 10.1559/152304003100011180.

American Meteorological Society. Enhancing weather information with probability forecasts. Information Statement of the American Meteorological Society, https://www.ametsoc.org/ams/index.cfm/about-ams/ams-statements/statements-of-the-ams-in-force/enhancing-weather-information-with-probability-forecasts/, May 2008.

Toni Amstad. *Wie verständlich sind unsere Zeitungen?* PhD thesis, University of Zurich, 1978.

Anthony D. Andre and Henry A. Cutler. Displaying uncertainty in advanced navigation systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 42(1):31–35, 1998. doi: 10.1177/154193129804200108.

Stavros Antifakos, Adrian Schwaninger, and Bernt Schiele. Evaluating the effects of displaying uncertainty in context-aware applications. In *UbiComp 2004: Ubiquitous Computing*, Lecture Notes in Computer Science, pages 54–69. Springer, Berlin/Heidelberg, 2004. ISBN 978-3-540-30119-6. doi: 10.1007/978-3-540-30119-6_4.

Aaron Bangor, Philip T. Kortum, and James T. Miller. An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6):574–594, 2008. doi: 10.1080/10447310802205776.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, October 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01.

Antonia Bauer and Ansbert Kneip. *Der große Wissenstest für Kinder - Was weißt du über die Welt?* Kiepenheuer & Witsch, Cologne, 2016. ISBN 978-3-462-31545-5.

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389–396, 2005. doi: 10.1037/1082-989X.10.4.389.

Ann M. Bisantz, Stephanie Schinzing Marsiglio, and Jessica Munch. Displaying uncertainty: Investigating the effects of display format and specificity. *Human Factors*, 47(4):777–796, 2005. doi: 10.1518/001872005775570916.

Ann M. Bisantz, Dapeng Cao, Michael Jenkins, Priyadarshini R. Pennathur, Michael Farry, Emilie Roth, Scott S. Potter, and Jonathan Pfautz. Comparing uncertainty visualizations for a dynamic decision-making task. *Journal of Cognitive Engineering and Decision Making*, 5(3):277–293, 2011. doi: 10.1177/1555343411415793.

Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. Falling asleep with Angry Birds, Facebook and Kindle: A large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56. ACM, New York, 2011. ISBN 978-1-4503-0541-9. doi: 10.1145/2037373.2037383.

Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, pages 3–27. Springer, London, 2014. ISBN 978-1-4471-6497-5. doi: 10.1007/978-1-4471-6497-5_1.

Nadia Boukhelifa and David John Duke. Uncertainty visualization: Why might it fail? In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, pages 4051–4056. ACM, New York, 2009. ISBN 978-1-60558-247-4. doi: 10.1145/1520340.1520616.

Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109. Springer, London, 2012. ISBN 978-1-4471-2804-5. doi: 10.1007/978-1-4471-2804-5_6.

David V. Budescu. Confidence in aggregation of opinions from multiple sources. In *Information Sampling and Adaptive Cognition*, pages 327–352. Cambridge University Press, New York, 2006. ISBN 0-521-53933-1.

David V. Budescu, Stephen Broomell, and Han-Hui Por. Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, 20(3):299–308, 2009. doi: 10.1111/j.1467-9280.2009.02284.x.

Stuart K. Card, Jock D. Mackinlay, and George G. Robertson. A morphological analysis of the design space of input devices. *ACM Transactions on Information Systems*, 9(2):99–122, April 1991. ISSN 1046-8188. doi: 10.1145/123078.128726.

Meredith A. Case, Holland A. Burwick, Kevin G. Volpp, and Mitesh S. Patel. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *The Journal of the American Medical Association*, 313(6): 625–626, February 2015. doi: 10.1001/jama.2014.17841.

Eun K. Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1143–1152. ACM, New York, 2014. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557372.

Edward T. Cokely, Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-Retamero. Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, 7(1):25–47, January 2012. ISSN 1930-2975. URL `http://journal.sjdm.org/11/11808/jdm11808.pdf`.

Leana Copeland and Tom Gedeon. 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, pages 285–288. ACM, New York, 2013. ISBN 978-1-4503-2525-7. doi: 10.1145/2541016.2541082.

Michael Correll and Michael Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, December 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346298.

Helen Couclelis. The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS*, 7(2):165–175, 2003. ISSN 1467-9671. doi: 10.1111/1467-9671.00138.

Matthew A. Cronin, Cleotilde Gonzalez, and John D. Sterman. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes*, 108(1):116–130, 2009. ISSN 0749-5978. doi: 10.1016/j.obhdp.2008.03.003.

Scott E. Crouter, Patrick L. Schneider, Murat Karabulut, and David R. Bassett. Validity of 10 electronic pedometers for measuring steps, distance, and energy cost. *Medicine & Science in Sports & Exercise*, 35(8):1455–1460, August 2003. ISSN 0195-9131. doi: 10.1249/01.mss.0000078932.61440.a2.

Allen Cypher and David C. Smith. KidSim: End user programming of simulations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 27–34. ACM Press/Addison-Wesley Publishing Co., New York, 1995. ISBN 0-201-84705-1. doi: 10.1145/223904.223908.

Alan Dix. Human-Computer Interaction. In *Encyclopedia of Database Systems*, pages 1327–1331. Springer US, Boston MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_192.

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5):377–398, May 2012. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2012.01.004.

Charles R. Ehlschlaeger, Ashton M. Shortridge, and Michael F. Goodchild. Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23 (4):387–395, 1997. ISSN 0098-3004. doi: 10.1016/S0098-3004(97)00005-8.

Stephen G. Eick. Data visualization sliders. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, UIST '94, pages 119–120. ACM, New York, 1994. ISBN 0-89791-657-3. doi: 10.1145/192426.192472.

Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, January 2002. ISSN 0028-0836. doi: 10.1038/415429a.

Jay Fenton and Kent Beck. Playground: An object-oriented simulation system with agent rules for children of all ages. *ACM SIGPLAN Notices*, 24(10): 123–137, September 1989. ISSN 0362-1340. doi: 10.1145/74878.74891.

Nivan Ferreira, Danyel Fisher, and Arnd C. Konig. 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, pages 571–580, 2014. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557131.

Kraig Finstad. The usability metric for user experience. *Interacting with Computers*, 22(5):323–327, September 2010. ISSN 0953-5438. doi: 10.1016/j.intcom.2010.04.004.

George W. Fitzmaurice. *Graspable user interfaces*. PhD thesis, University of Toronto, 1997.

Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32 (3):221–233, June 1948. doi: 10.1037/h0057532.

Jacqueline Frick and Christoph Hegg. Can end-users' flood management decision making be improved by information about forecast uncertainty? *Atmospheric Research*, 100(2):296–303, 2011. ISSN 0169-8095. doi: 10.1016/j.atmosres.2010.12.006.

Nahum Gershon. Visualization of an imperfect world. *IEEE Computer Graphics and Applications*, 18(4):43–45, July 1998. ISSN 0272-1716. doi: 10.1109/38.689662.

Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995. ISSN 1939-1471. doi: 10.1037/0033-295X.102.4.684.

Gerd Gigerenzer, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo, and Konstantinos V. Katsikopoulos. "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis*, 25(3):623–629, June 2005. ISSN 1539-6924. doi: 10.1111/j.1539-6924.2005.00608.x.

Tilmann Gneiting and Adrian E. Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, October 2005. ISSN 0036-8075. doi: 10.1126/science.1115255.

Miriam Greis. Entwicklung von Simulationswerkzeugen für Laien – Herausforderungen und Ziele. In *Digitale Welten: Neue Ansätze in der Wirtschafts- und Sozialkybernetik*, pages 91–105. Duncker & Humblot GmbH, Berlin, 2014.

Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. Investigating representation alternatives for communicating uncertainty to non-experts. In *Human-Computer Interaction – INTERACT 2015*, Lecture Notes in Computer Science, pages 256–263. Springer International Publishing, Cham, 2015. ISBN 978-3-319-22723-8. doi: 10.1007/978-3-319-22723-8_21.

Miriam Greis, Passant El.Agroudy, Hendrik Schuff, Tonja Machulla, and Albrecht Schmidt. 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, pages 52:1–52:4. ACM, New York, 2016. ISBN 978-1-4503-4763-1. doi: 10.1145/2971485.2971535.

Miriam Greis, Emre Avci, Albrecht Schmidt, and Tonja Machulla. 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, pages 828–840. ACM, New York, 2017a. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025998.

Miriam Greis, Tilman Dingler, Albrecht Schmidt, and Chris Schmandt. Leveraging user-made predictions to help understand personal behavior patterns. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, pages 104:1–104:8. ACM, New York, 2017b. ISBN 978-1-4503-5075-4. doi: 10.1145/3098279.3122147.

Miriam Greis, Jessica Hullman, Michael Correll, Matthew Kay, and Orit Shaer. 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, pages 593–600. ACM, New York, 2017c. ISBN 978-1-4503-4656-6. doi: 10.1145/3027063.3027091.

Miriam Greis, Hendrik Schuff, Marius Kleiner, Niels Henze, and Albrecht Schmidt. Input controls for entering uncertain data: Probability distribution sliders. *Proceedings of the ACM on Human-Computer Interaction*, 1(1): 3:1–3:17, June 2017d. ISSN 2573-0142. doi: 10.1145/3095805.

Theresia Gschwandtnei, Markus Bögl, Paolo Federico, and Silvia Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):539–548, January 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467752.

Fangfang Guo, Yu Li, Mohan S. Kankanhalli, and Michael S. Brown. An evaluation of wearable activity monitoring devices. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia*, PDM '13, pages 31–34. ACM, New York, 2013. ISBN 978-1-4503-2397-0. doi: 10.1145/2509352.2512882.

Tomislav Hengl and Norair Toomanian. Maps are not what they seem: Representing uncertainty in soil-property maps. In *Proceedings of Accuracy 2006*, pages 805–813. Institute Geogràfico Português, Lisbon, 2006. ISBN 972-8867-271. URL http://www.spatial-accuracy.org/system/files/Hengl2006accuracy.pdf.

Jürgen Hotz. *Duden - Testen Sie Ihre Allgemeinbildung*. Dudenverlag, Mannheim, 2013. ISBN 978-3-411-90914-8.

Jürgen Hotz. *Duden - Testen Sie Ihre Allgemeinbildung 2*. Dudenverlag, Berlin, 2014. ISBN 978-3-411-90913-1.

Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. Proactive response to eye movements. In *Human-Computer Interaction – INTERACT 2003*, INTERACT '03, pages 129–136. IOS press, Amsterdam, 2003. ISBN 1-58603-363-8.

Harald Ibrekk and M. Granger Morgan. Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, 7(4):519–529, 1987. ISSN 1539-6924. doi: 10.1111/j.1539-6924.1987.tb00488.x.

Christopher H. Jackson. Displaying uncertainty with shading. *The American Statistician*, 62(4):340–347, 2008. doi: 10.1198/000313008X370843.

Susan Joslyn and Sonia Savelli. Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, 17 (2):180–195, June 2010. ISSN 1469-8080. doi: 10.1002/met.190.

Susan L. Joslyn and Jared E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, 18(1):126–140, 2012. ISSN 1939-2192. doi: 10.1037/a0025185.

Susan L. Joslyn and Rebecca M. Nichols. Probability or frequency? Expressing forecast uncertainty in public weather forecasts. *Meteorological Applications*, 16(3):309–314, September 2009. ISSN 1469-8080. doi: 10.1002/met.121.

Susan L. Joslyn, Limor Nadav-Greenberg, Meng U. Taing, and Rebecca M. Nichols. The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Applied Cognitive Psychology*, 23(1):55–72, January 2009. ISSN 1099-0720. doi: 10.1002/acp.1449.

Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2201–2210. ACM, New York, 2015. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702479.

Jakob Karolus, Paweł W. Woźniak, Lewis L. Chuang, and Albrecht Schmidt. Robust gaze features for enabling language proficiency awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI

'17, pages 2998–3010. ACM, New York, 2017. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025601.

Matthew Kay, Dan Morris, mc schraefel, and Julie A. Kientz. 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, pages 401–410. ACM, New York, 2013. ISBN 978-1-4503-1770-2. doi: 10.1145/2493432.2493456.

Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. How good is 85%?: A survey tool to connect classifier evaluation to acceptability of accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 347–356. ACM, New York, 2015. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702603.

Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 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, pages 5092–5103. ACM, New York, 2016. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858558.

Sherman Kent. Words of estimative probability. *Journal of the American Intelligence Professional*, 8(4):49–65, 1964.

Anass Lasram, Sylvain Lefebvre, and Cyrille Damez. Scented sliders for procedural textures. In *Eurographics 2012 - Short Papers*, Eurographics '12. The Eurographics Association, Geneve, 2012. doi: 10.2312/conf/EG2012/short/045-048.

Averill M. Law. *Simulation Modeling and Analysis*. McGraw-Hill Education, New York, 5th edition, 2015. ISBN 987-1-259-25438-3.

Dan Ledger and Daniel McCaffrey. Inside Wearables: How the science of human behavior change offers the secret to long-term engagement. Blog: `https://blog.endeavour.partners/inside-wearable-how-the-science-of-human-behavior-change-offers-the-secret-to-long-term-engagement-a15b3c7d4cf3`, January 2014.

Wiliam Leiss. Three phases in the evolution of risk communication practice. *The ANNALS of the American Academy of Political and Social Science*, 545(1): 85–94, May 1996. ISSN 0002-7162. doi: 10.1177/0002716296545001009.

James R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995. doi: 10.1080/10447319509526110.

Brian Y. Lim and Anind K. Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 195–204. ACM, New York, 2009. ISBN 978-1-60558-431-7. doi: 10.1145/1620545.1620576.

Brian Y. Lim and Anind K. Dey. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 415–424. ACM, New York, 2011. ISBN 978-1-4503-0630-0. doi: 10.1145/2030112.2030168.

Isaac M. Lipkus. Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, 27(5):696–713, 2007. doi: 10.1177/0272989X07307271.

Isaac M. Lipkus, Greg Samsa, and Barbara K. Rimer. General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21(1):37–44, February 2001. doi: 10.1177/0272989X0102100105.

Alan M. MacEachren. Visualizing uncertain information. *Cartographic Perspective*, 13(13):10–19, November 1992. doi: 10.14714/CP13.1000.

Alan M. MacEachren, Anthony Robinson, Susan Hopper, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, July 2005. doi: 10.1559/1523040054738936.

Alan M. MacEachren, Robert E. Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2496–2505, December 2012. ISSN 1077-2626. doi: 10.1109/TVCG.2012.279.

Adrian Madsen, Adam Larson, Lester Loschky, and N. Sanjay Rebello. 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, pages 193–196. ACM, New York, 2012. ISBN 978-1-4503-1221-9. doi: 10.1145/2168556.2168591.

Dawn M. Marsh-Richard, Erin S. Hatzis, Charles W. Mathias, Nicholas Venditti, and Donald M. Dougherty. Adaptive Visual Analog Scales (AVAS): A modifiable software program for the creation, administration, and scoring of visual analog scales. *Behavior Research Methods*, 41(1):99–106, February 2009. ISSN 1554-3528. doi: 10.3758/BRM.41.1.99.

Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5421–5432. ACM, New York, 2016. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858063.

Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 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, pages 4000–4004. ACM, New York, 2016. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858247.

Betty Hearn Morrow, Julie L. Demuth, and Jeffrey K. Lazo. Communicating weather forecast uncertainty: An exploratory study with broadcast meteorologists. Final report of focus groups conducted at the 36th AMS conference on broadcast meteorology, SocResearch Miami and National Center for Atmospheric Research, 2008.

Rebecca E. Morss, Julie L. Demuth, and Jeffrey K. Lazo. Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Weather and Forecasting*, 23(5):974–991, October 2008. ISSN 0882-8156. doi: 10.1175/2008WAF2007088.1.

Rebecca E. Morss, Jeffrey K. Lazo, and Julie L. Demuth. Examining the use of weather forecasts in decision scenarios: Results from a US survey with implications for uncertainty communication. *Meteorological Applications*, 17 (2):149–162, June 2010. ISSN 1469-8080. doi: 10.1002/met.196.

Allan H. Murphy and Robert L. Winkler. Forecasters and probability forecasts: The responses to a questionnaire. *Bulletin of the American Meteorological Society*, 52(3):158–166, March 1971. doi: 10.1175/1520-0477(1971)052<0158:FAPFTR>2.0.CO;2.

Allan H. Murphy, Sarah Lichtenstein, Baruch Fischhoff, and Robert L. Winkler. Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society*, 61(7):695–701, July 1980. doi: 10.1175/1520-0477(1980)061<0695:MOPPF>2.0.CO;2.

Limor Nadav-Greenberg and Susan L. Joslyn. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, September 2009. doi: 10.1518/155534309X474460.

National Research Council. *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. The National Academies Press, Washington D.C., 2006. ISBN 978-0-309-10255-1. doi: 10.17226/11699.

John P. Norton, James D. Brown, and Jaroslav Mysiak. To what extent, and how, might uncertainty be defined? Comments engendered by "Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support". *Integrated Assessment Journal*, 6(1):83–88, 2006. ISSN 1389-5176. URL `http://journals.sfu.ca/int_assess/index.php/iaj/article/view/9/195`.

Christopher Olston and Jock D. Mackinlay. Visualizing data with bounded uncertainty. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '02, pages 37–40. IEEE Computer Society, Washington D.C., 2002. ISBN 0-7695-1751-X. doi: 10.1109/INFVIS.2002.1173145.

Masakazu Osada, Holmes Liao, and Ben Shneiderman. Alpha slider: Searching textual lists with sliders. Department of Computer Science Technical Report, University of Maryland, April 1993.

Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, and Eeva Raita. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, 16(1):105–114, January 2012. ISSN 1617-4917. doi: 10.1007/s00779-011-0412-2.

Alex T. Pang, Craig M. Wittenbrink, and Suresh K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, November 1997. ISSN 1432-2315. doi: 10.1007/s003710050111.

Florian Pappenberger, Elisabeth Stephens, Jutta Thielen, Peter Salamon, David Demeritt, Schalk Jan van Andel, Fredrik Wetterhall, and Lorenzo Alfieri. Visualizing probabilistic flood forecast information: Expert preferences and perceptions of best practice in uncertainty communication. *Hydrological Processes*, 27(1):132–146, January 2013. ISSN 1099-1085. doi: 10.1002/hyp.9253.

Heike Pfersdorff and Iris Glahn. *Duden - Testen Sie Ihr Wissen! Das Allgemeinbildungsquiz*. Dudenverlag, Berlin, 2015. ISBN 978-3-411-91104-2.

Kristin Potter, Joe Kniss, Richard Riesenfeld, and Chris R. Johnson. Visualizing summary statistics and uncertainty. *Computer Graphics Forum*, 29(3):823–832, June 2010. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2009.01677.x.

Kristin Potter, Paul Rosen, and Chris R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, volume 377 of *IFIP Advances in Information and Communication Technology*, pages 226–249. Springer, Berlin/Heidelberg, 2012. ISBN 978-3-642-32677-6. doi: 10.1007/978-3-642-32677-6_15.

Steven F. Railsback and Volker Grimm. *Agent-based and individual-based modeling: A practical introduction*. Princeton University, Princeton NJ, 2012. ISBN 978-0-691-13673-8.

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

Majken K. Rasmussen, Esben W. Pedersen, Marianne G. Petersen, and Kasper Hornbæk. Shape-changing interfaces: A review of the design space and open research questions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 735–744. ACM, New York, 2012. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2207781.

Mitchel Resnick. StarLogo: An environment for decentralized modeling and decentralized thinking. In *Conference Companion on Human Factors in Computing Systems*, CHI '96, pages 11–12. ACM, New York, 1996. ISBN 0-89791-832-0. doi: 10.1145/257089.257095.

Gordan Ristovski, Tobias Preusser, Horst K. Hahn, and Lars Linsen. Uncertainty in medical visualization: Towards a taxonomy. *Computers & Graphics*, 39: 60–73, April 2014. ISSN 0097-8493. doi: 10.1016/j.cag.2013.10.015.

Maria Riveiro. Evaluation of uncertainty visualization techniques for information fusion. In *Proceedings of the 10th International Conference on Information Fusion*, Fusion '07, pages 1–8. IEEE Computer Society, Washington D.C., 2007. doi: 10.1109/ICIF.2007.4408049.

Anne Roudaut, Abhijit Karnik, Markus Löchtefeld, and Sriram Subramanian. Morphees: Toward high "shape resolution" in self-actuated flexible mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 593–602. ACM, New York, 2013. ISBN

978-1-4503-1899-0. doi: 10.1145/2470654.2470738. URL http://doi.acm.org/10.1145/2470654.2470738.

Mark S. Roulston and Todd R. Kaplan. A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts. *Meteorological Applications*, 16(2):237–244, June 2009. ISSN 1469-8080. doi: 10.1002/met.113.

Mark S. Roulston, Gary E. Bolton, Andrew N. Kleit, and Addison L. Sears-Collins. A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting*, 21(1):116–122, February 2006. ISSN 0882-8156. doi: 10.1175/WAF887.1.

Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 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, pages 247–250. ACM, New York, 2006. ISBN 1-59593-390-5. doi: 10.1145/1152215.1152267.

Dominik Sacha, Hansi Senaratne, Bum C. Kwon, Geoffrey Ellis, and Daniel A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, January 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467591.

Jibonananda Sanyal, Song Zhang, Gargi Bhattacharya, Phil Amburn, and Robert J. Moorhead. A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1209–1218, November 2009. ISSN 1077-2626. doi: 10.1109/TVCG.2009.114.

Jeffer E. Sasaki, Amanda Hickey, Marianna Mavilia, Jacquelynne Tedesco, Dinesh John, Sarah K. Keadle, and Patty S. Freedson. Validation of the fitbit wireless activity tracker for prediction of energy expenditure. *Journal of Physical Activity and Health*, 12(2):149–154, February 2015. doi: 10.1123/jpah.2012-0495.

Sonia Savelli and Susan Joslyn. The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, 27(4):527–541, July/August 2013. ISSN 1099-0720. doi: 10.1002/acp.2932.

Julia Schwarz, Scott Hudson, Jennifer Mankoff, and Andrew D. Wilson. 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, pages 47–56. ACM, New York, 2010. ISBN 978-1-4503-0271-5. doi: 10.1145/1866029.1866039.

Wilfried Seibicke. *Duden - Wie schreibt man gutes Deutsch? Eine Stilfibel.* Dudenverlag, Mannheim, 1969. ISBN 3-411-01137-8.

Orit Shaer, Oded Nov, Johanna Okerlund, Martina Balestra, Elizabeth Stowell, Lauren Westendorf, Christina Pollalis, Jasmine Davis, Liliana Westort, and Madeleine Ball. 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, pages 661–672. ACM, New York, 2016. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858397.

Orit Shaer, Oded Nov, Lauren Westendorf, and Madeleine Ball. Communicating personal genomic information to non-experts: A new frontier for Human-Computer Interaction. *Foundations and Trends® in Human-Computer Interaction*, 11(1):1–62, May 2017. ISSN 1551-3955. doi: 10.1561/1100000067.

Robert E. Shannon. Introduction to the art and science of simulation. In *Proceedings of the 30th Conference on Winter Simulation*, WSC '98, pages 7–14. IEEE Computer Society Press, Los Alamitos CA, 1998. ISBN 0-7803-5134-7.

Meredith Skeels, Bongshin Lee, Greg Smith, and George G. Robertson. Revealing uncertainty for information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '08. ACM, New York, 2008. ISBN 978-1-60558-141-5. doi: 10.1145/1385569.1385637.

David C. Smith, Allen Cypher, and Jim Spohrer. KidSim: Programming agents without a programming language. *Communications of the ACM*, 37(7):54–67, July 1994. ISSN 0001-0782. doi: 10.1145/176789.176795.

Randall B. Smith. The alternate reality kit: An animated environment for creating interactive simulations. In *Proceedings of the 1986 IEEE Computer Society Workshop on Visual Languages*, pages 99–106. IEEE Computer Society Press, Silver Spring MD, 1986.

David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, September 2011. ISSN 0036-8075. doi: 10.1126/science.1191181.

Linda B. Sweeney and John D. Sterman. Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review*, 16(4):249–286, Winter 2000. ISSN 1099-1727. doi: 10.1002/sdr.198.

Susanne Tak and Alexander Toet. Color and uncertainty: It is not always black and white. In *EuroVis - Short Papers*, EuroVis '14, pages 55–59. The Eurographics Association, Geneve, 2014. ISBN 978-3-905674-69-9. doi: 10.2312/eurovisshort.20141157.

Barry N. Taylor and Chris E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST Technical Note 1297, National Institute of Standards and Technology, 1994.

John Taylor. *Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements, 2nd Edition*. University Science Books, New York, 1997. ISBN 0-935702-42-3.

Karl Halvor Teigen and Magne Jørgensen. When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, 19(4):455–475, May 2005. ISSN 1099-0720. doi: 10.1002/acp.1085.

Jakob Tholander and Stina Nylander. 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, pages 2913–2922. ACM, New York, 2015. ISBN 978-1-4503-3145-6. doi: 10.1145/2702123.2702482.

Judi Thomson, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. A typology for visualizing uncertainty. In *Proceedings of the IS&TSPIE's Symposium on Electronic Imaging*, volume 5669 of *IS&TSPIE '05*, pages 146–157, 2005. doi: 10.1117/12.587254.

Seth Tisue and Uri Wilensky. Netlogo: A simple environment for modeling complexity. Technical report, New England Complex Systems Institute, 2004.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. In *Utility, Probability, and Human Decision Making*, volume 11 of *Theory and Decision Library*, pages 141–162. Springer Netherlands, Dordrecht, 1975. ISBN 978-94-010-1834-0. doi: 10.1007/978-94-010-1834-0_8.

Tina C. Walber, Ansgar Scherp, and Steffen Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2065–2074. ACM, New York, 2014. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557025.

Warren E. Walker, Poul Harremoës, Jan Rotmans, Jeroen P. van der Sluijs, Marjolein B. A. van Asselt, Peter Janssen, and Martin P. Krayer von Krauss. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4(1):5–17, 2003. doi: 10.1076/iaij.4.1.5.16466.

Gudrun Wallentin and Adrijana Car. A framework for uncertainty assessment in simulation models. *International Journal of Geographical Information Science*, 27(2):408–422, 2013. doi: 10.1080/13658816.2012.715163.

Thomas S. Wallsten, David V. Budescu, Amnon Rapoport, Rami Zwick, and Barbara Forsyth. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115(4):348–365, December 1986. ISSN 1939-2222. doi: 10.1037/0096-3445.115.4.348.

Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, November 2007. ISSN 1077-2626. doi: 10.1109/TVCG.2007.70589.

Craig M. Wittenbrink, Elijah Saxon, Jeff J. Furman, Alex T. Pang, and Suresh K. Lodha. Glyphs for visualizing uncertainty in environmental vector fields. In *Proceedings of the IS&TSPIE's Symposium on Electronic Imaging*, IS&TSPIE '96, pages 266–279, 1996. doi: 10.1117/12.205940.

Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 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, pages 143–146. ACM, New York, 2011. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1978963.

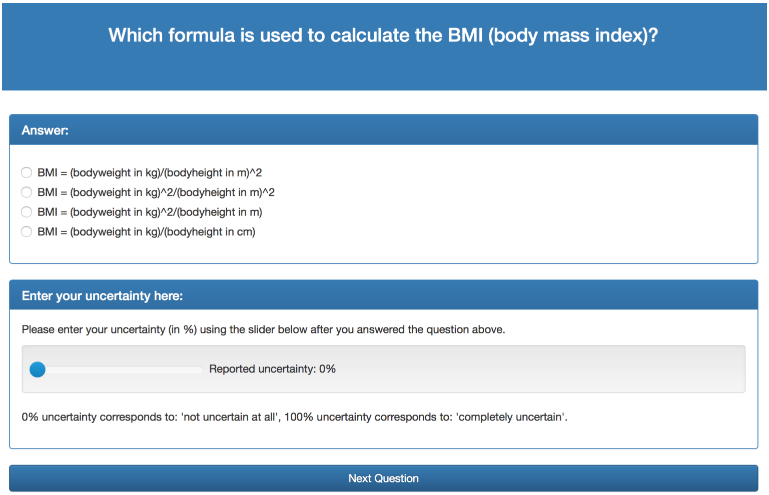
VI

Appendices

Appendix A

Study Interface for Enhancing Standard Input Controls

In this appendix, we include screenshots of our web-based study interface for enhancing standard input controls described in Section 6.1.



Which formula is used to calculate the BMI (body mass index)?

Answer:

○ BMI = (bodyweight in kg)/(bodyheight in m)^2
○ BMI = (bodyweight in kg)^2/(bodyheight in m)^2
○ BMI = (bodyweight in kg)^2/(bodyheight in m)
○ BMI = (bodyweight in kg)/(bodyheight in cm)

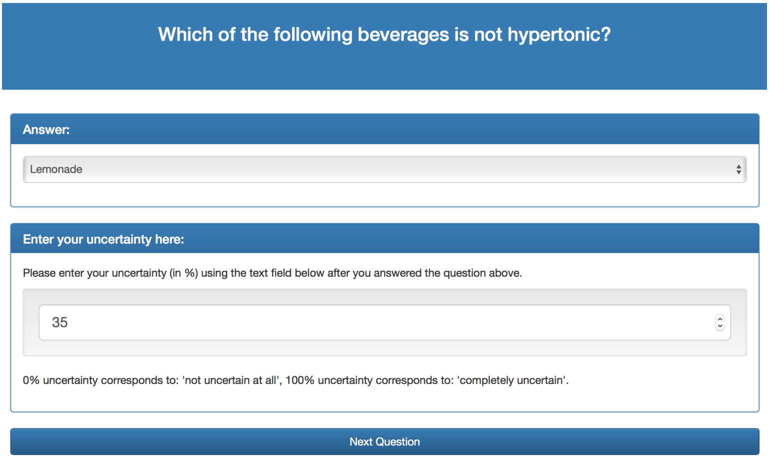
Enter your uncertainty here:

Please enter your uncertainty (in %) using the slider below after you answered the question above.

Reported uncertainty: 0%

0% uncertainty corresponds to: 'not uncertain at all', 100% uncertainty corresponds to: 'completely uncertain'.

(a) Study interface 1: Radio buttons as main input method and a numeric slider to report uncertainty as percentage.



Which of the following beverages is not hypertonic?

Answer:

Lemonade

Enter your uncertainty here:

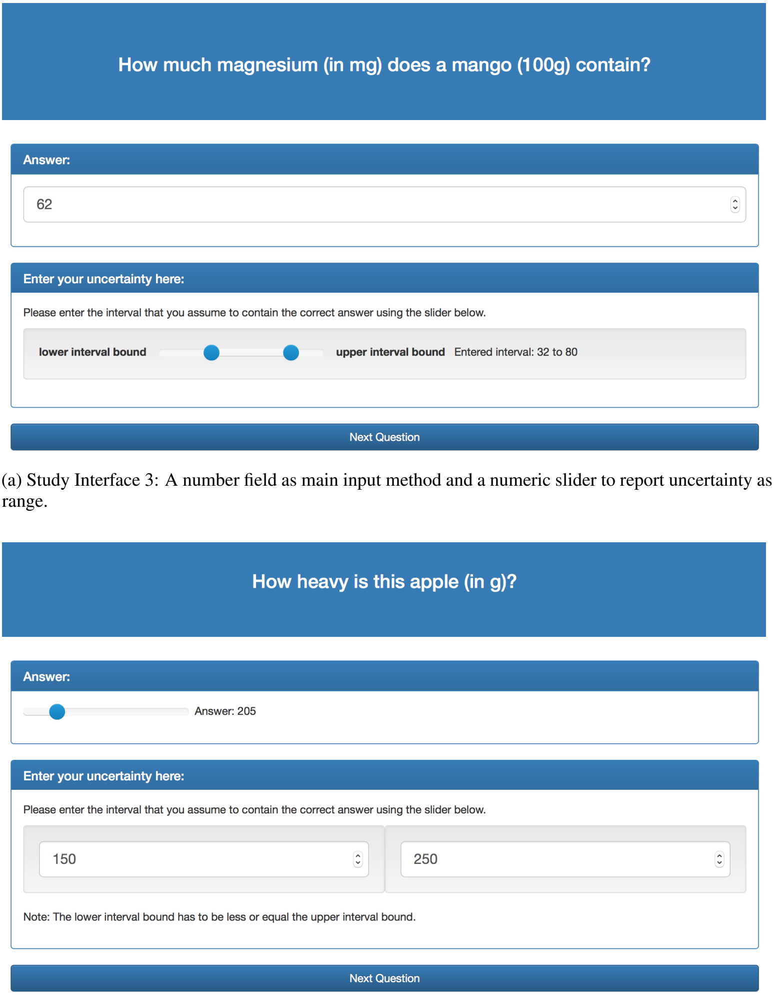
Please enter your uncertainty (in %) using the text field below after you answered the question above.

35

0% uncertainty corresponds to: 'not uncertain at all', 100% uncertainty corresponds to: 'completely uncertain'.

(b) Study interface 2: Drop down menu as main input method and a number field to report uncertainty as percentage.

Figure A.1: Web-based study interface with the two different main input methods and uncertainty input methods for mutiple-choice questions.



How much magnesium (in mg) does a mango (100g) contain?

Answer:

62

Enter your uncertainty here:

Please enter the interval that you assume to contain the correct answer using the slider below.

lower interval bound upper interval bound Entered interval: 32 to 80

(a) Study Interface 3: A number field as main input method and a numeric slider to report uncertainty as range.

How heavy is this apple (in g)?

Answer:

Answer: 205

Enter your uncertainty here:

Please enter the interval that you assume to contain the correct answer using the slider below.

150

250

Note: The lower interval bound has to be less or equal the upper interval bound.

(b) Study Interface 4: A slider with two thumbs as main input method and a pair of number fields to report uncertainty as range.

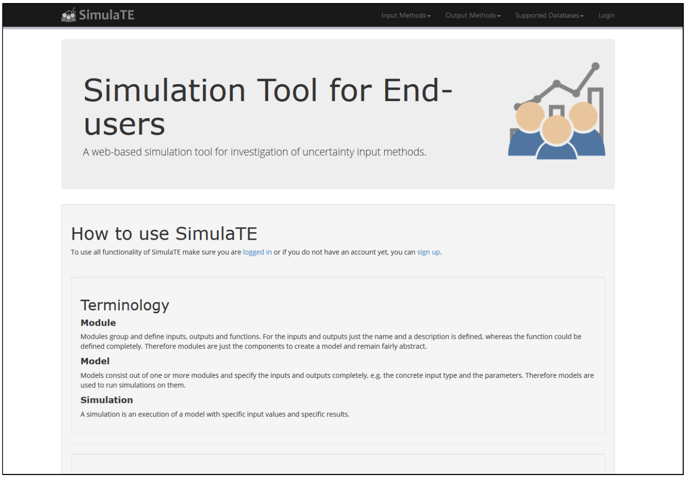
Figure A.2: Web-based study interface with the two different main input methods and uncertainty input methods for numerical questions.

220A Study Interface for Enhancing Standard Input Controls

Appendix B

Screenshots of SimulaTE

In this appendix, we include screenshots of our web-based simulation tool SimulaTE described in Chapter 9.



Simulation Tool for End-users

A web-based simulation tool for investigation of uncertainty input methods.

How to use SimulaTE

To use all functionality of SimulaTE make sure you are logged in or if you do not have an account yet, you can sign up.

Terminology

Module

Modules group and define inputs, outputs and functions. For the inputs and outputs just the name and a description is defined, whereas the function could be defined completely. Therefore modules are just the components to create a model and remain fairly abstract.

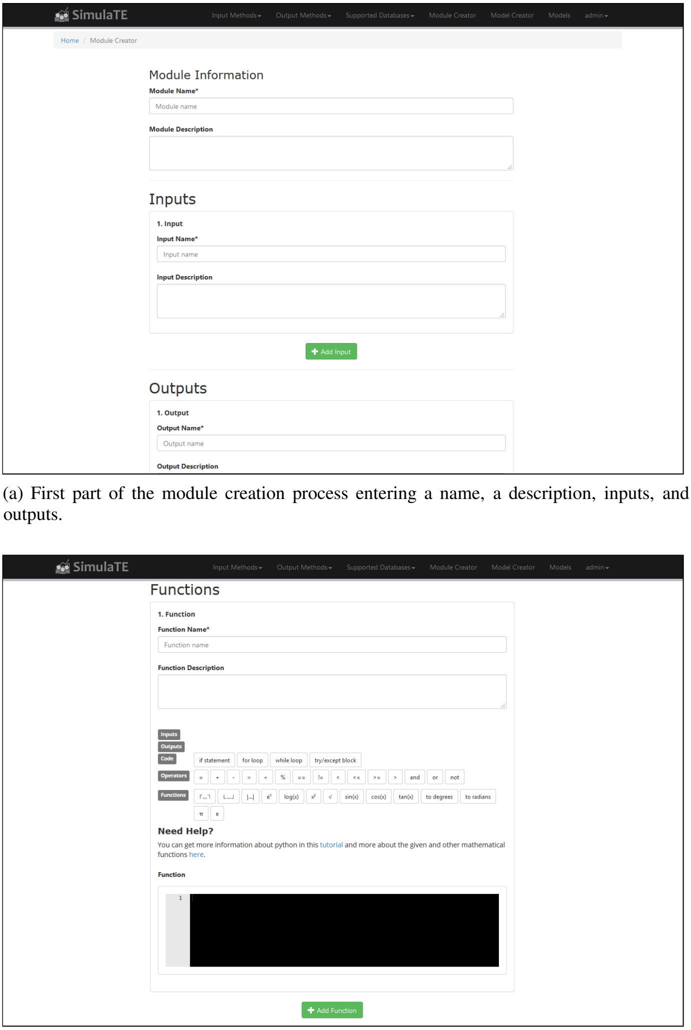
Model

Models consist out of one or more modules and specify the inputs and outputs completely, e.g. the concrete input type and the parameters. Therefore models are used to run simulations on them.

Simulation

A simulation is an execution of a model with specific input values and specific results.

Figure B.1: Introduction page of SimulaTE with an explanation of terms and functionality of the tool.

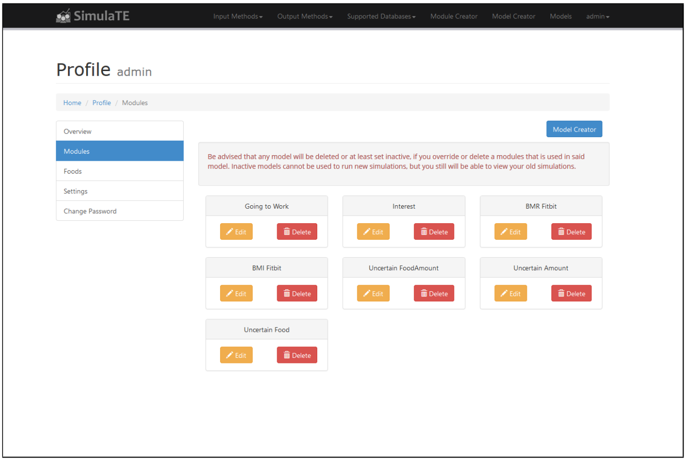


(a) First part of the module creation process entering a name, a description, inputs, and outputs.

You can get more information about python in this tutorial and more about the given and other mathematical functions here.

(b) Second part of the module creation process entering functions.

Figure B.2: The module creator of SimulaTE allows the users to create new modules. A module consists of a name, a description, inputs, outputs, and functions.

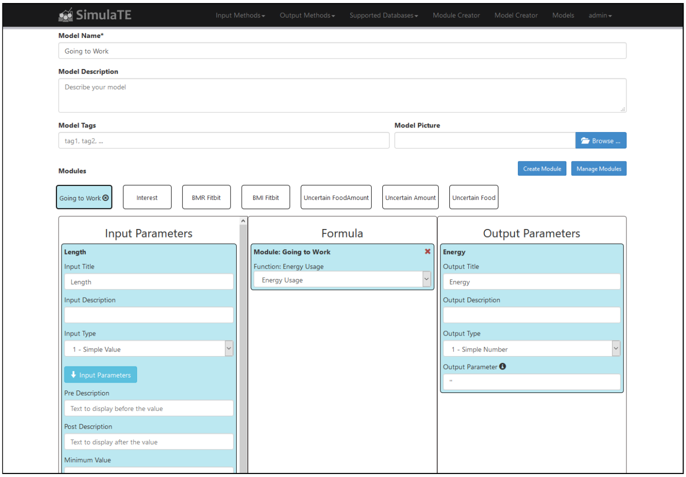


Profile admin

Be advised that any model will be deleted or at least set inactive, if you override or delete a modules that is used in said model. Inactive models cannot be used to run new simulations, but you still will be able to view your old simulations.

Going to Work	Interest	BMR Fitbit
BMI Fitbit	Uncertain FoodAmount	Uncertain Amount
Uncertain Food		

Figure B.3: The module overview of SimulaTE allows the users to edit or delete their modules.



Model Name*: Going to Work

Model Description: Describe your model

Model Tags: tag1, tag2, ...

Model Picture

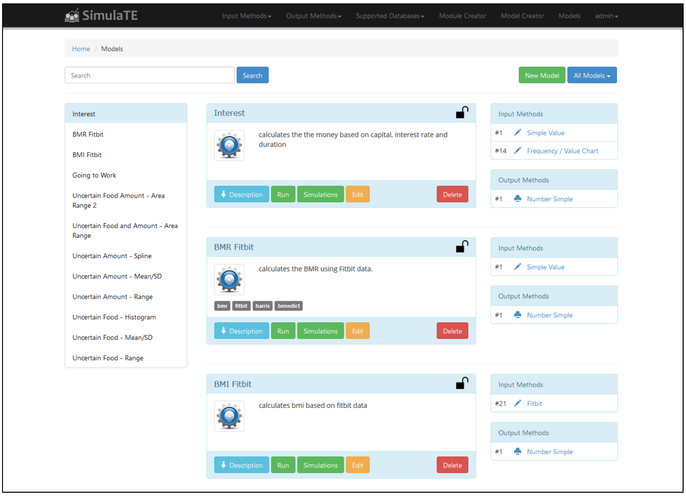
Modules

Input Parameters
- Length
- Input Title: Length
- Input Description
- Input Type: 1 - Simple Value
- Pre Description: Text to display before the value
- Post Description: Text to display after the value
- Minimum Value

Formula
- Module: Going to Work
- Function: Energy Usage
- Energy Usage

Output Parameters
- Energy
- Output Title: Energy
- Output Description
- Output Type: 1 - Simple Number
- Output Parameter: ..

Figure B.4: The model creator allows the users to create a model. The model has a name, a description, tags and a picture. The users can select specific input and output methods and a formula from the modules which can be selected by dragging them to the editor.



Interest
BMR Fitbit
BMI Fitbit
Going to Work
Uncertain Food Amount - Area Range 2
Uncertain Food and Amount - Area Range
Uncertain Amount - Spline
Uncertain Amount - Mean/SD
Uncertain Amount - Range
Uncertain Food - Histogram
Uncertain Food - Mean/SD
Uncertain Food - Range

Interest

calculates the the money based on capital, interest rate and duration

Input Methods
#1 Simple Value
#14 Frequency / Value Chart

Output Methods
#1 Number Simple

BMR Fitbit

calculates the BMR using Fitbit data.

Input Methods
#1 Simple Value

Output Methods
#1 Number Simple

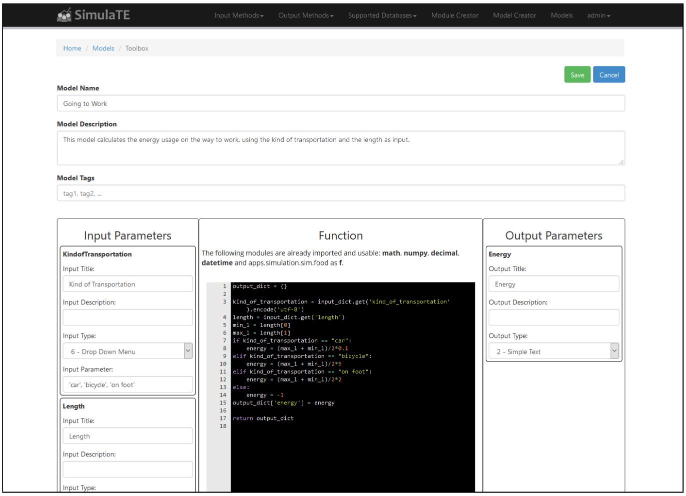
BMI Fitbit

calculates bmi based on fitbit data

Input Methods
#21 Fitbit

Output Methods
#1 Number Simple

Figure B.5: The model overview provides users with the possible to search for models and the possible to edit, delete, or run a model.



Model Name
Going to Work

Model Description
This model calculates the energy usage on the way to work, using the kind of transportation and the length as input.

Model Tags
tag1, tag2, ...

Input Parameters

KindofTransportation
Input Title: Kind of Transportation
Input Description:
Input Type: 6 - Drop Down Menu
Input Parameter: 'car', 'bicycle', 'on foot'

Length
Input Title: Length
Input Description:
Input Type:

Function

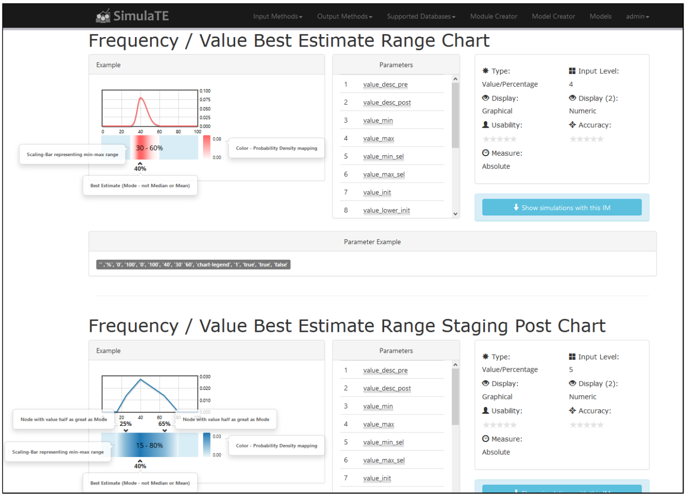
The following modules are already imported and usable: **math**, **numpy**, **decimal**, **datetime** and apps.simulation.sim.food as **f**.

```
output_dict = {}

kind_of_transportation = input_dict.get('kind_of_transportation'
    ).encode('utf-8')
length = input_dict.get('length')
min_l = length[0]
max_l = length[1]
if kind_of_transportation == "car":
    energy = (max_l + min_l)/2*0.1
elif kind_of_transportation == "bicycle":
    energy = (max_l + min_l)/2*5
elif kind_of_transportation == "on foot":
    energy = (max_l + min_l)/2*2
else:
    energy = -1
output_dict['energy'] = energy

return output_dict

```

Output Parameters

Energy
Output Title: Energy
Output Description:
Output Type: 2 - Simple Text

Figure B.6: The model editor allows users to adapt the inputs, outputs, and functions of a model.


SimulaTE
Input Methods
Output Methods
Supported Databases
Module Creator
Model Creator
Models
admin
0.100
0.075
0.050
0.025
0.000
0
20
40
60
80
100
30 - 60%
0.08
0.00
40%
Scaling-Bar representing min-max range
Color - Probability Density mapping
Best Estimate (Mode - not Median or Mean)
Show simulations with this IM
0.030
0.020
0.010
0.000
25%
65%
Node with value half as great as Mode
Node with value half as great as Mode
15 - 80%
0.03
0.00
40%
Scaling-Bar representing min-max range
Color - Probability Density mapping
Best Estimate (Mode - not Median or Mean)

Frequency / Value Best Estimate Range Chart

Example

Parameters

1 value_desc_pre
2 value_desc_post
3 value_min
4 value_max
5 value_min_sel
6 value_max_sel
7 value_init
8 value_lower_init

Type: Value/Percentage
Input Level: 4
Display: Graphical
Display (2): Numeric
Usability: ★★★★★
Accuracy: ★★★★★
Measure: Absolute

Parameter Example

`'' ,'%', '0', '100', '0', '100', '40', '30' '60', 'chart-legend', '1', 'true', 'true', 'false'`

Frequency / Value Best Estimate Range Staging Post Chart

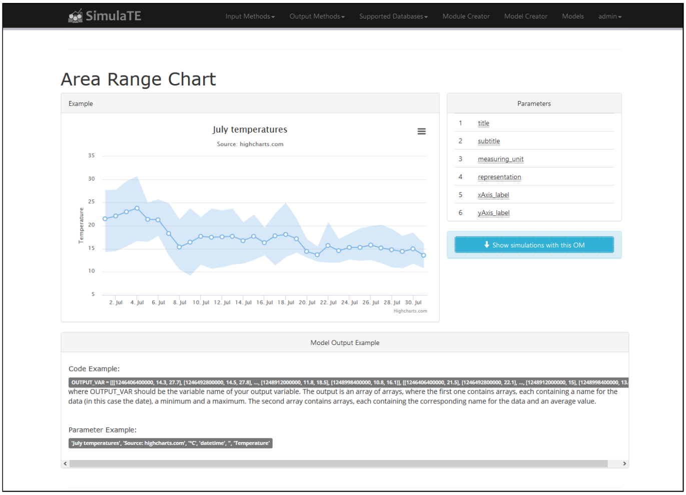
Example

Parameters

1 value_desc_pre
2 value_desc_post
3 value_min
4 value_max
5 value_min_sel
6 value_max_sel
7 value_init

Type: Value/Percentage
Input Level: 5
Display: Graphical
Display (2): Numeric
Usability: ★★★★★
Accuracy: ★★★★★
Measure: Absolute

Figure B.7: The overview of the input methods shows users all available input methods with associated parameters. Additionally, a classification of the input methods and an example for valid parameters are shown.



Area Range Chart

Example

Parameters

1 title
2 subtitle
3 measuring_unit
4 representation
5 xAxis_label
6 yAxis_label

Model Output Example

Code Example:

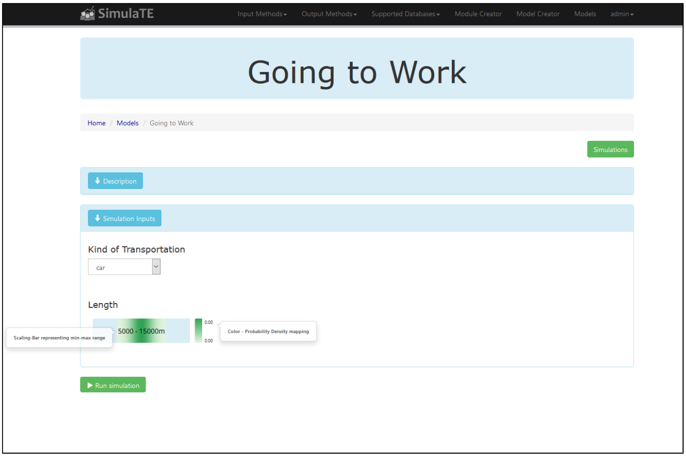
```
OUTPUT_VAR = [[[1246406400000, 14.3, 27.7], [1246492800000, 14.5, 27.8], ..., [1248912000000, 11.8, 18.5], [1248998400000, 10.8, 16.1]], [[1246406400000, 21.5], [1246492800000, 22.1], ..., [1248912000000, 15], [1248998400000, 13.
```

where OUTPUT_VAR should be the variable name of your output variable. The output is an array of arrays, where the first one contains arrays, each containing a name for the data (in this case the date), a minimum and a maximum. The second array contains arrays, each containing the corresponding name for the data and an average value.

Parameter Example:

`'July temperatures', 'Source: highcharts.com', '°C', 'datetime', '', 'Temperature'`

Figure B.8: The overview of the output methods shows users all available output methods with associated parameters, for example the area range chart. Code and parameter examples are included.



Going to Work

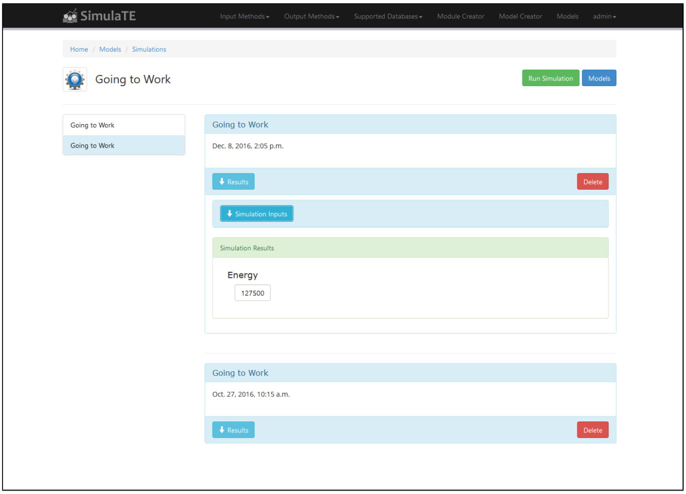
Kind of Transportation

car

Length

5000 - 15000m

Figure B.9: User interface for running a simulation in SimulaTE.



Going to Work

Going to Work

Dec. 8, 2016, 2:05 p.m.

Energy

127500

Going to Work

Oct. 27, 2016, 10:15 a.m.

Figure B.10: Overview on simulation results in SimulaTE.

Miriam Greis

A Systematic Exploration of Uncertainty in Interactive Systems

Uncertainty is an inherent part of our everyday life. Humans have to deal with uncertainty every time they make a decision. The importance of uncertainty additionally increases in the digital world. Machine learning and predictive algorithms introduce statistical uncertainty to digital information. In addition, the rising number of sensors in our surroundings increases the amount of statistically uncertain data, as sensor data is prone to measurement errors. Hence, there is an emergent need for practitioners and researchers in Human-Computer Interaction to explore new concepts and develop interactive systems able to handle uncertainty. Such systems should not only support users in entering uncertainty in their input, but additionally present uncertainty in a comprehensible way.

The main contribution of this thesis is the exploration of the role of uncertainty in interactive systems and how novel input and output methods can support researchers and designers to efficiently and clearly communicate uncertainty. By using empirical methods of Human-Computer Interaction and a systematic approach, we present novel input and output methods that support the comprehensive communication of uncertainty in interactive systems. We further integrate our results in a simulation tool for end-users.

Based on related work, we create a systematic overview of sources of uncertainty in interactive systems to support the quantification of uncertainty and identify relevant research areas. The overview can help practitioners and researchers to identify uncertainty in interactive systems and either reduce or communicate it. We then introduce new concepts for the input of uncertain data. We enhance standard input controls, develop specific slider controls and tangible input controls, and collect physiological measurements. We also compare different representations for the output of uncertainty to make recommendations for their usage. Furthermore, we analyze how humans interpret uncertain data und make suggestions on how to avoid misinterpretation and statistically wrong judgements. We embed the insights gained from the results of this thesis in an end-user simulation tool to make it available for future research. The tool is intended to be a starting point for future research on uncertainty in interactive systems and foster communicating uncertainty and building trust in the system. Overall, our work shows that user interfaces can be enhanced to effectively support users with the input and output of statistically uncertain information.