INAUGURAL-DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich-Mathematischen Gesamtfakultät

der

Ruprecht-Karls-Universität

Heidelberg

vorgelegt von

Rumyana Proynova

aus Sofia

Tag der mündlichen Prüfung: .....

Measuring anticipated satisfaction

Betreut von: Prof. Dr. Barbara Paech
Prof. Dr. Thomas Wetter

Abstract

When developing a software system, one of the early steps is to create a requirements specification. Validating this specification saves implementation effort which might be otherwise spent on building a system with the wrong features. Ideally, this validation should involve many stakeholders representing different groups, to ensure coverage of a variety of viewpoints. However, the usual requirements validation methods such as personal interviews only allow the involvement of a few stakeholders before the costs become prohibitive, so it is difficult to apply them at the needed scale.

If the finished software system contains undesirable features, they are likely to be discovered during usability testing. Many usability methods can involve a high number of users at a low cost, for example satisfaction surveys and A/B testing in production. They can give high quality information about improving the system, but they require a completed system or at least an advanced prototype before they can be used.

We create a method for measuring user satisfaction before building the system, which we call *anticipated satisfaction* to distinguish it from the *actual satisfaction* measured after the user has experienced the system. The method uses a questionnaire which contains short descriptions of the software system's features, and asks the users to imagine how satisfied they would be when using a system with the described features. The method is flexible, as we do not create a single questionnaire to use. Instead, we give guidance on which variables can be measured with the questionnaire, and how to create questions for them. This allows the development team to tailor the questionnaire to the specific situation in their project. When we applied it in two validation studies, it discovered significant issues and was rated favorably by both the software development team and the users.

Our method contributes to the discipline of software engineering by offering a new option for validating software requirements. It is more scalable than interviewing users, and can be employed before the implementation phase, allowing for early problem detection. The effort required to apply it is low, and the information gained is seen as useful by both developers and managers, which makes it a good candidate for use in commercial projects.

Zusammenfassung

Beim Entwickeln eines Softwaresystems ist einer der ersten Schritte das Erstellen einer Anforderungsspezifikation. Das Validieren dieser Spezifikation reduziert den Implementierungsaufwand, der möglicherweise beim Entwickeln eines Systems mit den falschen Features entstanden wäre. Im Idealfall werden viele Stakeholder in diese Validierung miteinbezogen, die unterschiedliche betroffene Gruppen repräsentieren, um das Abdecken unterschiedlicher Perspektiven zu gewährleisten. Allerdings erlauben die traditionellen Validierungsmethoden wie Interviews nur das Involvieren von wenigen Stakeholdern, bevor die Kosten untragbar werden, so dass es schwierig ist, diese Methoden im benötigten Umfang anzuwenden.

Falls die fertige Software unerwünschte Features enthält, kann man diese mit hoher Wahrscheinlichkeit bei einem Usability Test entdecken. Viele Usability Methoden erlauben das Involvieren vieler Benutzer bei niedrigen Kosten, wie zum Beispiel Zufriedenheitsbefragungen und A/B Tests in der Produktionsphase. Sie produzieren hochqualitative Information zum Verbessern des Softwaresystems, aber sie benötigen ein fertiggestelltes System oder zumindest ein fortgeschrittenes Prototyp um angewendet zu werden.

Wir erstellen eine Methode für das Messen von Benutzerzufriedenheit bevor das System implementiert ist, und nennen diese Metrik “erwartete Zufriedenheit”, um sie von der “eigentlichen Zufriedenheit” zu unterscheiden, die erst gemessen werden kann nachdem der Benutzer Erfahrungen mit dem System gesammelt hat. Die Methode benutzt einen Fragebogen, der kurze Beschreibungen von Softwarefeatures enthält, und bittet die Benutzer sich vorzustellen, wie zufrieden sie mit einem System mit den beschriebenen Features wären. Die Methode ist flexibel, da wir keinen festen Fragebogen vorgeben. Stattdessen geben wir eine Anleitung, welche Variablen mit dem Fragebogen gemessen werden können, und wie man Fragen zu diesen Variablen formuliert. Das erlaubt es dem Entwlicklungsteam, den Fragebogen an der spezifischen Situation in ihrem Projekt anzupassen. Als wir die Methode in zwei Validierungsstudien anwendeten, entdeckten wir wichtige Probleme, und die Methode wurde positiv vom Entwicklunsteam und von den teilnehmenden Benutzern angenommen.

Unsere Methode trägt zur Disziplin des Software Engineering bei, indem sie eine neue Option für das Validieren von Softwareanforderungen bietet. Sie ist besser skalierbar als Benutzerinterviews, und kann vor der Implementationsphase angewendet werden, so dass sie die frühe Entdeckung von Problemen erlaubt. Der Aufwand für die Anwendung ist niedrig, und die produzierte Information wird sowohl von den Entwicklern als auch von den Führungskräften als nützlich angesehen, deswegen ist sie die Methode für die Anwendung in kommerziellen Projekten gut geeignet.

Acknowledgements

Writing this dissertation was a long journey. My first thanks go to my supervisor Prof. Barbara Paech, who was there at every step of it. She made sure that I did things the right way, taught me how to work scientifically, and never lost faith in me even in the most difficult of times. Many thanks also to Prof. Wetter, who always brought in new and interesting angles to all discussions we had, and gave me many new ideas. I also thank the faculty of Heidelberg University for accepting me as a doctoral student and giving me support during the process.

I spent the last five years working at the dissertation while holding a job at the German Cancer Research Center (DKFZ). There, I have to thank my group leader, Dr. Claudia Galuschka, who always showed great support for me, no matter what I was going through, and created for me the opportunity to conduct the central validation study for this thesis. After I changed divisions, the division head, Prof. Frank Ückert was also supportive of this work and helped me improve the way I present it to others.

Both at Heidelberg University and the DKFZ, my colleagues have been involved in the work in some way. I thank Tom-Michael Hesse, who created a prototype needed for one of my user studies, Benjamin Roth, who slipped in the role of a questionnaire creater, and Nico Helfrich, who implemented an invitation for the users to partake in my study right into his software application. I wrote papers together with Sven Koch and Andreas Wicht, and that was always a fruitful collaboration. No less important than these tangible contributions was the time spent in informal discussions and just the knowledge of having friendly coworkers who make it enjoyable to go to the office every day. At the university of Heidelberg, these were Alexander Delater, Robert Heinrich, Hanna Remmel, Ulli Abelein, Thorsten Merten, Gabrielle Zorn-Pauli, Paul Hübner, Christian Kücherer, Anja Kleebaum, Leonore Dietrich, Marcus Seiler and Thomas Quirchmayr. When I had a technical issue, Willi Springer took it off my hands.

At work, there were Jürgen Trauth, who always knew how to do something in C#, and Karin Müller-Decker, with whom I gathered most of my experience in requirements elicitation and who helped me understand the dissertation process from a different point of view. And there is of course my current team, who have been amazing at working with me I while was busy learning to be a manager and finishing a dissertation at the same time. Paul Weingardt, Saher Maqsood, Max Ataian, Pururava Mishra and Naile Büber, thank you for being a great team.

I wouldn't have been able to go through all this without my friends. Manfred Klingmann patiently taught me how to paddle a kayak, and was always there for a short excursion on the Neckar or a weeklong whitewater adventure in the Alps. Volker Gärtner taught me to face my fears and was always there to bring me to solid ground when I capsized. Udo Mattern, Wolfgang Lerbs and Silvia Weiss opened their doors for me and gave me support when I needed it most. And Jolene Brown, Wren Middleton and Stephanie Heldmayer have been a constant presence in my life, physically far away, but always there when I needed somebody to talk.

And of course, there is my family. They rooted for me from thousands of kilometers away and waited until my schedule permitted a rare visit. Without them and their sacrifices, I wouldn't be here writing this. Mom, dad, everybody, thank you from the bottom of my heart!

Contents

Part I

Preliminaries

1 Introduction

In software engineering, satisfaction measurement is a form of validation, a way of answering the question “Did we build the right system?”. Satisfaction measurement is a very thorough way of validation, but it is only possible to conduct in the very late stages of a project, after an alpha version of the system has been delivered to users. As the cost of introducing changes grows disproportionally when the project stages advance, it is desirable to also use validation at earlier project stages.

One of the earliest artefacts available in a software development project is a requirements specification. Requirements validation is a major part of the requirements engineering process, and is defined as checking the consistency, completeness, and accuracy of the requirements specification [118]. Textbooks suggest multiple techniques for requirements validation [156, 161, 118, 7], which can be classified as pre-reviews, reviews, prototyping, model-based, testing-based and viewpoint-oriented [163].

Empirical research has found requirements validation to be a major success factor in software projects [89]. Software companies use the techniques described above [89, 163], although they rarely follow the exact textbook descriptions, preferring to adapt them for their own use [89]. Disadvantages include the high effort [163] and the lack of know-how in teams [89]. Nevertheless, requirements validation techniques are reported to constitute between 3.1% and 10% of total project effort, and successful teams spend a higher proportion of project time on requirements engineering activities than unsuccessful teams [89, 163].

Not all of the validation methods described above involve users. For those that do, the participation is typically limited to a few key users, due to the time intensive nature of the methods (e.g. having a requirements engineer sit together with the user for a walkthrough) and the complexity of explaining the process to the user. In the worst case, these users are the same who were interviewed for eliciting the requirements. This can confirm that their opinions were understood and documented correctly, but it does not give any information to how prevalent these opinions are among the broader user population. There have been recorded cases where system failure was traced back to tailoring the requirements to the preferences of a renowned domain expert who turned out to not be representative of the majority of users [124].

We propose to employ user satisfaction measurement as a new technique for requirements validation. Its advantages lie in the ability to use it with a large number of users, being directly tied to system success [51], and being based on an intuitive concept which is understood and valued

by both users and product owners. It is also a good way to increase user participation, which has been recognized as a further success factor for software development projects [2]. This would make satisfaction measurement a valuable tool for requirements validation, complementing existing techniques and extending the options available to requirements engineers.

1.1 Goal and contributions

Our goal is to investigate the potential of capturing users' prediction of their satisfaction with a future system as an approach for early validation. We term the satisfaction measured at this stage *anticipated satisfaction*, as opposed to the *actual satisfaction* which is measured after exposure to a finished system. To achieve our goal, we develop a method for measuring anticipated satisfaction and show its usefulness when applied for early validation of a requirements specification.

We divide our research project in several stages, each of which results in a distinct contribution. First, we conduct a systematic literature review, which produces a *list of concepts related to satisfaction*. We also create a *categorization schema*, ASMA, which we use to analyze our findings.

The list is a resource required for applying the method we develop later. At the same time, it is a result which adds to existing theoretical knowledge in the field of usability, enabling better understanding of the factors involved in establishing satisfaction.

For the second result, we create a *model of anticipated satisfaction*. It is validated in two empirical studies. This model provides a basis for understanding and future research of anticipated satisfaction.

The studies not only confirm our theory, but we also use them for testing the *practical feasibility* of eliciting anticipated satisfaction. They show that users are willing and capable of answering the type of questions needed for measuring anticipated satisfaction. These questions have high cognitive complexity, requiring the users to reason about their future feelings about a subject which they cannot experience directly. The discovery that these questions are not overwhelming, and that users do not refuse answering them for other reasons, is an important finding. Without this, we would not have been able to apply the theory.

Based on these results, we create the central contribution of this thesis. We develop MUSA, a *method for measuring anticipated user satisfaction*. It involves using a questionnaire to show users features descriptions based on the requirements specification and asking them to describe what they think of a system with these features. The questions are chosen from a list of satisfaction related concepts. The results can be used to improve the specification. The method was validated in a series of studies in a real-world development project.

With these contributions, our research has both theoretical and practical implications. It contributes to the field of usability by synthesizing important results in satisfaction measurement

and providing a model of anticipated satisfaction. In the area of requirements engineering, it suggests a novel approach to specification validation and provides a method which implements that approach. For practitioners, it provides a method whose application can reduce work effort and uncertainty, and was designed to have a low adoption barrier.

The broader context for our method is a software development project. Due to the diversity of software development, it is unlikely that a single method will fit with any development process and any type of software system which can be developed. We therefore narrow down the context for which the method is developed and validated. We assume 1) a greenfield development 2) of a system with which users interact 3) to fulfill a task 4) related to one of their information needs. Such systems include business information systems, medical information systems and other types of software systems in professional use. In 1), we exclude extending or upgrading existing systems. While it is conceivable that the method will work for that too, we do not explore how previous knowledge about an existing system interacts with anticipating satisfaction of new features, so cannot tell how applying the method in this context differs from the greenfield situation. With 2), we define that we do not consider systems with which the user does not interact directly, such as embedded systems. The third condition rules out systems which are not used to complete a task that has meaning outside of the system, such as video games. The fourth condition selects systems which create, find or process data, as opposed to e.g. automation systems which control machines.

The above limitations were chosen to provide maximum flexibility for our method. They allow us to focus on a large and well-studied class of software systems for which there are standard methods for measuring actual satisfaction. This makes those systems a prime target for our research, while producing a manageable scope for the basic method. It is conceivable that in future work, it can be extended to cover systems which are not considered in this dissertation.

1.2 Structure of the thesis

The structure of the thesis reflects the research stages and contributions listed in the last section. An overview of the structure is graphically represented in figure 1.1.

The current chapter forms the preliminary part together with the background chapter. The background chapter briefly reviews the methods currently in use for satisfaction measurement, introduces the theoretical constructs we use in our research, and lists quality goals for the satisfaction measurement method.

The second part describes the bulk of our research and its results. It starts with chapter 3, which describes the concept of anticipated satisfaction and creates a model for it. In that chapter, we analyze a simple pre-exposure measurement of satisfaction-related concepts, and focus on the additional factors which change the results opposed to a post-exposure measurement of the same

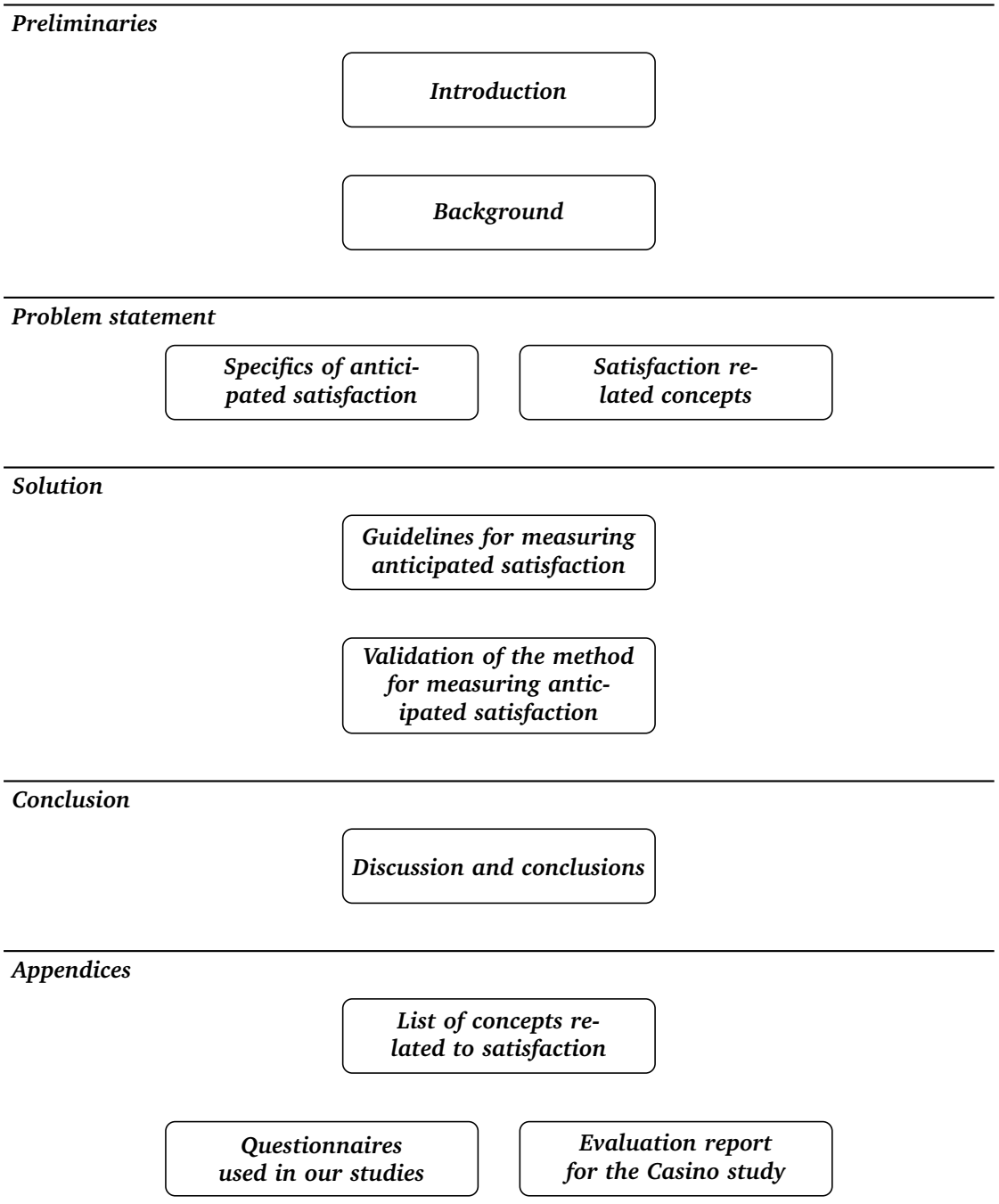


Figure 1.1: Structure of the thesis

concepts. We use the results of two empirical study to create and refine a model of anticipated satisfaction.

Chapter 3 shows that classic satisfaction-related concepts are needed for the measurement of anticipated satisfaction. In chapter 4, we conduct a systematic literature review on satisfaction-related concepts. It provides us with a list of concepts which can be used to measure anticipated satisfaction, and we use metaanalysis to determine how well these concepts correlate with satisfaction.

The knowledge from the first two chapters allows us to construct a method for measuring anticipated satisfaction. We describe this method in chapter 5, which gives step-by-step instructions for practitioners who wish to employ the method in their development project. It describes the preparation, data gathering and evaluation of the data needed to find out the level of anticipated satisfaction among potential users.

To validate our method, we use two empirical studies, described in chapter 6. For the first study, we apply our own method with users who have not had access to the system, then measure actual satisfaction with the same system. This allows us to compare anticipated and actual satisfaction. In the second study, a person not affiliated with our project applied the method for his project, giving us insight in the practical applicability of the method.

The last part of the thesis is the summary. In it we discuss the potential use of the method and list potential research areas made available through its existence.

The appendices contain a short characterization of each satisfaction-related concept we found in the systematic literature review, a reprint of the questionnaires we used in our empirical studies, and the evaluative report which was provided to the product owner in the second validation study. The raw data and the source code for the analysis scripts is not included in the thesis, but will be made available on request.

2 Background

While anticipated satisfaction has not been studied in depth before, measuring actual satisfaction is a well-developed area, and we base our research on its findings. It is necessary to understand the principles behind satisfaction measurement in general before exploring a method for measuring anticipated satisfaction. This chapter provides the needed background knowledge, summarizing the basics of satisfaction measurement and explaining conventions, assumptions and terms we use in the rest of the thesis.

We start with a section describing the different views of what satisfaction is, and introduce the definition we employ in our research. The following section presents the methods typically employed for satisfaction measurement. Section 2.3 introduces *theory*, *idealized model* and *data model* – three constructs which are central to analysing existing satisfaction measurement methods and creating the new method for anticipated satisfaction. Then we briefly define the usage of the term *metric*, which has a specific meaning in the domain of software engineering [69]. The chapter ends with a section setting up the expectations for our new method and describing quality criteria for evaluating it.

2.1 Defining satisfaction

Satisfaction is a concept which permeates human life. It is highly visible in popular culture, media, religious beliefs and several fields of science. Scientists study people's satisfaction with their job [140], their life [53], or even the experience of being a hospital patient [190].

The word *satisfaction* has multiple closely related dictionary meanings, but many of its meanings are very specific and do not apply to our research. We are not concerned with algorithms which can satisfy given conditions, or facts which can be proven to a court's satisfaction. We use the primary meaning, in which satisfaction has a sentient subject and requires an object with which the subject is satisfied. While there is some variety regarding satisfaction in software engineering – for example, there is literature on software developers' satisfaction with their jobs [72] – our research focuses strictly on a situation where a *user* experiences satisfaction with a *software system*. [1].

[1] Consequently, we use the terms *satisfaction* and *user satisfaction* interchangeably, and in the context of this thesis even the unqualified form *satisfaction* is taken to mean *user satisfaction* unless specifically stated otherwise.

Even for this specific meaning, definitions of *satisfaction* are highly varied. Since there is no single established authoritative definition, we extracted the common themes from existing definitions in literature to arrive at a new one which combines the known characteristics of satisfaction and emphasizes the nuances in meaning which are most pertinent to our analysis in later chapters.

Our full definition of is

> User satisfaction is a user's affective evaluation of a software system.

In this definition, user satisfaction is an affective construct. This is the prevailing view in the literature both on user satisfaction and other types of satisfaction, supported for example in [20, 9, 181, 19, 153], and is more common than alternatives which represent user satisfaction as an attitude [68] or a belief [100]. We also consider it to be evaluative in nature – that is, the user judges the system to be satisfying or not [9, 185, 206]. While we define satisfaction to be about a software system, we do not make assumptions about the situation which causes and forms the feeling of satisfaction. It may be the use of the system, as some authors claim in their definitions [26, 27, 68], or another event such as reflecting on the system, discussing it with colleagues, or something else. This level of detail about the user's cognitive and affective processes is beyond the scope of this thesis. We only make a pragmatic distinction relevant to researchers or practitioners who wish to use the method we present in this thesis. We define the level of satisfaction experienced after use to be *actual satisfaction*, while any level reported prior to use is considered *anticipated satisfaction*. This allows us to compare the two and investigate the suitability of pre-exposure satisfaction measurements to arrive at conclusions about the software's usability.

This definition of satisfaction has several consequences for measurement. First, it makes satisfaction subjective, or, in terms of measurement theory, satisfaction has to be measured with *pragmatic measurements* as opposed to *representational measurements* [78]. There is no single "real value" of satisfaction to be captured the way a physical object has a real mass which physicists can measure. Rather, each user has his or her own level of satisfaction, which is as valid as the satisfaction level of any other user. Any satisfaction measurement for a given software system has to be an aggregate measurement representative of a group of users, ideally the whole target population. Satisfaction cannot be measured to arbitrarily high precision, because at some point the measurement error becomes insignificantly small in comparison to individual differences between users.

The second consequence is that satisfaction measurement requires users. As satisfaction is subjective, it cannot be derived from the system itself or from other information. In the terms of measurement theory, this makes satisfaction an *external attribute* rather than an *internal attribute* of the software system [62]. This does not mean that satisfaction is completely arbitrary, because with everything else being equal, a better quality system has a higher chance of creating a high satisfaction response in the user. System quality and other internal attributes of the system strongly influence the level of satisfaction, but they do not create it deterministically.

Defining satisfaction as an internal attribute of software has been attempted in some contexts. The International communication union defines user satisfaction to be equivalent to quality of service, and quality of service is defined in terms of technical parameters such as packet loss and echo [99]. While this may be a reasonable proxy for a system with minimal interaction (the user does not interact with the transport layer of a telephone system directly), it does not measure the same construct as the one used in our research. In our work, satisfaction is always the user's response to a system.

Lastly, as satisfaction is an emotion, its measurement is limited to approaches which can measure an individual's affective state. Emotions are "complex, largely automated programs of actions" [49]. They include multiple physiological reactions not controllable by the individual, such as activation of brain nuclei, the secretion of hormones (e.g. cortisol in the case of fear) and muscular contraction and relaxation, which for some emotions result in signature face expressions [59]. The individual can feel his or her emotions, and is motivated by them to engage in certain behaviors. These four aspects of emotion – neurophysiological changes, changes in musculature, behavioral changes and subjective awareness of the emotion – lead to four possible approaches for measuring emotion.

One approach is to directly observe the neurophysiology of emotions. This requires an imaging technology such as functional magnetic resonance imaging (fMRI) and is not suitable for widespread adoption in software engineering. The second approach involves registration of the physiological reactions caused by the emotion. This approach is sometimes employed in human-computer interaction (HCI) studies [166], but requires expensive equipment and good technical skills. We are not aware of any research showing its validity when the user is imagining rather than experiencing the system. A third approach would be to observe the individual's behavior and infer an emotion from it. In our method, we cannot observe the user during interaction with the system, and direct observation of behavior outside of system interaction would be expensive and intrusive (e.g. following the user's communication and counting instances of recommending the system to others). Thus we do not use observation of user behavior.

In our research, we choose the fourth approach, known as *self-reporting*. It relies on the user feeling an emotion and reporting the feeling, typically in a questionnaire. This approach is scalable, accepted by users, and has a long tradition in measuring HCI concepts, especially satisfaction [179]. This is also evidenced in the next section, which gives an overview of how the field of measuring user satisfaction developed over the years.

2.2 History of user satisfaction measurement

The first widespread instrument for measuring user satisfaction was published in 1983 by Bailey and Pearson [11]. They extracted potential satisfaction-related concepts from earlier studies and discussed their list with information system experts and managers. The result was a list of

39 concepts which can be measured in a questionnaire. Later research created and validated shorter questionnaires based on their list, such as the Ives instrument [100] and the Baroudi and Orlikowski instrument [16]. A later influential instrument was published in 1988, the Doll and Torkzadeh EUCS (end-user computing satisfaction) instrument [55]. It used factor analysis to extract five satisfaction related concepts, to be elicitated with a 12-item questionnaire.

In the 1990s, research on user satisfaction measurement was further refined, with authors addressing criticism on early instruments. The focus was shifted from satisfaction as the goal variable to choosing a different, more business-relevant goal variable, with satisfaction being its major driver. The most prominent examples are DeLone and McLean's information system success measurement [52], and Davis' Technology Acceptance Model (TAM) [50]. Both were introduced at the beginning of the decade and underwent considerable development and validation, to be revised in the early 2000s. DeLone and McLean's model was made available in a revisited version [51], while the TAM merged with other theories and was expanded into the UTAUT (unified theory of acceptance and use of technology) by Venkatesh [184], which was again revised in 2008 [183]. Later, theories of technology acceptance pedigree were combined with satisfaction theory in the work of Wixom and Todd [195].

Most of these instruments were published not in software engineering venues, but rather in journals about management, which traditionally belong to the field of economics and business administration. Despite this positioning, they developed independently from the research on *customer satisfaction*, which is the subject of business administration, specifically marketing. While the *users* of a software system are not necessarily the *customers* paying for a software product or service, the two groups frequently overlap. Also, researchers with marketing background can define their field as *consumer satisfaction*, which focuses on the person consuming the good, rather than the one paying for it. The most influential research in that area is represented by Oliver's work [153], which produced the expectation-disconfirmation theory of satisfaction. This theory was introduced in software user satisfaction measurement in 2001 by Bhattacherjee [20] and has gained popularity in the field. In our systematic literature review, 18 out of 136 primary studies measured disconfirmation, a concept not present in instruments derived from Bailey and Pearson's work.

From the mid- to late 2000s on, there have been no new major theories on user satisfaction. Rather, research has focused on applying existing theory and instruments in new contexts, such as measuring the satisfaction of users with Internet banking [25] or mobile phones [60]. Another trend is to explore the contribution of other theories or concepts to satisfaction, such as trust [108], aesthetics [40] or subjective norm [130].

In computer science literature, user satisfaction is defined as one of the three factors of usability [98], the other two being effectiveness and efficiency. Thus it is mostly seen in a usability context, and measuring satisfaction is traditionally part of measuring usability. In practice, little distinction is made between the measurement of related concepts like *satisfaction*, *usability* or *user experience*. While there have been some efforts to differentiate between them [125], the

terms are still inconsistently used and have much overlap. The measurement methods employed for usability and user experience are the same, and both include satisfaction measurements.

From this background, a multitude of usability and satisfaction measurement instruments have developed. They tend to have less stringent theoretical backing and are geared towards easy application in commercial settings. Sometimes they are created by software companies rather than academic researchers. Some notable representatives of questionnaire-based usability instruments include IsoMetrics [70], PSSUQ [129], QUIS [39], SUMI [111], SUS [24] and WAMMI [110].

Due to this historical development, today's researchers and practitioners are faced with a large number of alternative questionnaires. Their use is very similar. The questionnaire is administered to participants who have used the system, the answers to each question are collected and aggregated in some way to arrive at a measurement of satisfaction. The difference between questionnaires is in the questions asked, and in the suggested evaluation methods for the gathered data.

2.3 Scientific theory and scientific models

In this thesis, we create a new questionnaire for satisfaction, based on the theory and models used in earlier satisfaction measurement studies. In this section, we provide a short introduction to the concepts of *model* and *theory*, as philosophy of science explains them.

A scientific theory is a "systematic explanatory scheme" [175]. It consists of rules which guide the behavior of a system [2]. In some scientific areas, these rules can be very exact, such as axioms, theorems and laws of nature. However, humans are not trivial machines [66], and a system which includes humans does not react the same way to equivalent input. Thus science branches which explain such systems, such as cognitive and social sciences, have more relaxed rules.

A *scientific model* is a description of a system in which a given theory applies. There is a large body of literature that analyzes models and classifies them in different categories [192, 122, 71]. A brief review of the field can be found in [114]. In our work, we follow the classification proposed in the Stanford encyclopedia of philosophy [67]. It distinguishes three main type of models – representational models of phenomena (with several subtypes), representational models of data, and models of theory. Two of these three model types, the *idealized representational model of phenomena* and the *representational model of data*, are relevant to our work, and are explained here in detail. For brevity, we refer to them as *idealized model* and *data model* respectively.

A data model is the result of an empirical study. The study measures variables in a system present

[2] The term *system* does not refer to a software system here, but is a concept used in theory of science. It is defined as "A portion of the universe that has been chosen for studying the changes that take place within it in response to varying conditions" [58]. To prevent confusion with the software system, in sections where both terms are used, we will call the subject of a theory the *system of interest*. In this section, we continue to call it simply *system*, as a software system is not discussed here.

Scientific construct	Theory	Idealized representational model	Data model
Description	A set of rules which govern a system's behavior	An abstract description of a system in which a theory applies	A fit of raw data to equations describing a system
Focused on	– rules about systems – system behavior (describing and predicting it)	– abstract concepts which describe entities from the real world – relationships between these concepts	– operationalized variables – measurements – calculation of parameters based on measured data
Example from physics [149]	Newton's theory of gravity	A sphere in free fall towards Earth's surface. The sphere and Earth are entities, while the distance covered by the sphere and its speed are concepts.	Measuring the time t an object falls a known distance d, then using the formula $d = \frac{gt^2}{2}$ to find a value for the standard gravity g
Example from satisfaction measurement	The theoretical background of the technology acceptance model [183]	Technology acceptance model (TAM)	Conducting a study measuring TAM concepts, and statistically determining the extent to which they are related.

Table 2.1: A comparison of *theory*, *idealized representational model* and *data model* as used in this thesis.

in the real world. The measurement results are then fitted to equations or formulas derived from the theory behind the study, to arrive at the final evaluation. This fitting can be as complex as doing multiple regression, or as simple as averaging the measured numbers to arrive at the final result. The exact mathematical or statistical method used depends on the chosen theory.

A real-world system with human elements cannot be replicated exactly. However, it is possible to create or find naturally occurring systems which have the same structure. The *idealized model* is an abstract description of the structure of a system. It characterizes the entities which comprise the system, the concepts describing these entities, and the relationships between the concepts. For example, in satisfaction measurement, a *user* is a real world entity, the user's *trust* in the software system is a concept, and "users with higher levels of trust are more satisfied with a software system" is a relationship between the concepts *trust* and *satisfaction*. If an experiment of the same design is conducted multiple times, this results in multiple different data models, but only one idealized representational model.

Table 2.1 provides a comparison of the three terms central to this section. It uses two examples to illustrate them. The first example comes from high school physics and should be understandable for readers who are not familiar with the field of satisfaction measurement. The second example shows how the three terms apply to constructs used in the scientific measurement of user satisfaction.

The three scientific constructs stand in close relationship to each other. The relationships in the three pairs of constructs are defined as follows:

Idealized model and theory An idealized model represents a system. The behavior of the system can be described with one or more theories. We will only consider simple cases, in which there is exactly one theory for each idealized model. Then the concepts of the idealized model will be the subject of the theory's rules.

Data model and theory A data model also represents a system. It is a quantitative model, and can be described in equations. A theory of the system determines which equations are expected to be valid in the data model, and the measurements are fitted to these equations. The variables of the data model will correspond to the subjects of the system's rules.

Idealized model and data model The system of a data model is an instance of the more abstract system described by the idealized model. Each variable in the data model is an operationalization of a concept of the idealized model. We will only consider cases in which a data model is an instance of exactly one idealized model.

When given a data model, it is always possible to construct a trivial idealized model by assigning a concept to each variable of the data model[3]. Similarly, when given an idealized model, it is always possible to construct a minimal theory by creating a rule which involves the relationships of the idealized model, although it will be lacking in explanatory mechanisms.

Due to these close relationships, researchers frequently do not make distinctions when reporting their research. For example, the background chapter of an article describing an empirical study can mix the structure of an idealized model with the theory rules applicable to the concepts, without explicitly noting them. Also the data model and the idealized model to which it corresponds are frequently treated as a single construct. We make the distinction where needed, but also frequently use shorter phrases when ambiguity is not a concern.

[3] We should note that there is no consistent usage of these terms in literature. In primary research, there is usually a 1:1 correspondence between a concept and a variable, so researchers frequently speak of *variables* only, even when they refer to the concepts being measured with these variables. A textbook on secondary research [45] uses the terms *construct* and *operation* for our *construct* and *variable* respectively.

2.4 Metrics in software engineering

"Metric" is a term commonly used in software engineering literature. Its popularity can be traced back to the work of Basili and Rombach, who introduced the Goal/Question/Metric approach in software engineering [17]. The IEEE Standard glossary of software engineering terminology defines it as

> a quantitative measure of the degree to which a system, component, or process possesses a given attribute.

In software engineering, the object of a metric is frequently a software system itself. This type of metric is called a *software metric*. While its usage has been subjected to criticism [69], mostly because it is not used in this sense in other fields, it is widely used in software engineering practice and theory [63].

Some concepts that describe software are complex and cannot be measured directly, so they are measured through intermediary concepts. For example, usability is frequently measured through concepts such as task success or number of errors [179]. When a concept is used to measure usability, it becomes a *usability metric*.

In this thesis, we describe the measuring of satisfaction through measurements of other concepts. In parallel to the established usage for usability, we use the term *satisfaction metric* to denote any concept which is measured in order to arrive at a measurement of satisfaction. This definition assumes that a satisfaction measurement is taking place. When we discuss the same concepts in a more general context, we use the term *satisfaction related concept* instead. It means that the concept is known to have some relationship to satisfaction and could potentially be used as a satisfaction metric.

2.5 Summary

In this chapter, we set up the background of the thesis. We provided a definition for satisfaction, and briefly sketched how it is measured. As the available satisfaction measurement methods are based on a model derived from a theory, we described what theories are and how they are related to models. This knowledge allows us to construct a new method, based on a new model. We also listed some qualities of a good satisfaction measurement method, against which we can evaluate the newly constructed method. This concludes the preliminary part of the thesis. The topic of the next part is the actual creation of the satisfaction measurement method.

Part II

Developing a method for measuring anticipated satisfaction

3 Specifics of anticipated satisfaction

The goal of our research is to design a method for measuring anticipated satisfaction, using methods for measuring actual satisfaction as the starting point. Similar to those, our method will be based on a questionnaire, to be filled out by the users. However, the differences between anticipated and actual satisfaction mean that existing questionnaires cannot be simply reused. Instead, our method will provide guidelines for creating a questionnaire.

The formulation of questions in a questionnaire can have a large impact on the answers [177]. Therefore, the guidelines should not only define the content of the question, but also recommend a structure known to produce the desired results. The cognitive demand placed on the users is higher than with questions for actual satisfaction, as they are required to reason about their hypothetical emotional state as opposed to describing their existing one. This high cognitive demand creates a risk of either the users denying to fill out any answers, or of their predictions' accuracy being too low for practical use. Thus our model should both define the content of the questions and recommend a question structure known to produce answers of sufficient quality.

In a questionnaire for actual satisfaction, the users base their answers on experience gained from using the system. Our method replaces that with knowledge about the system that is derived from reading about its requirements. The users cannot be expected to read the complete requirements specification, as it is vast and many parts of it are too technical for them to understand. Instead, our method calls for using a simpler representation of the requirements. It will be included in the questionnaire, and the questions will reference it directly. Before we create the method, we have to determine which requirements can be included, and what is a suitable way to present them to the users.

This chapter provides the groundwork for our method by defining requirements representations and questions which can be used in its questionnaire. It has two goals:

- Goal 1: To gather first experiences with instruments of anticipated satisfaction.
- Goal 2: To create an idealized model of anticipated satisfaction.

To achieve these goals, we start with an exploratory study. It measures anticipated satisfaction, together with other metrics which can be used to validate and better understand this measurement. We describe this study in section 3.1. The results of this study are used to construct a model of anticipated satisfaction in section 3.2. The model is then validated with a second study,

described in section 3.3. The model can be used to determine the concepts which should be measured in a questionnaire for anticipated satisfaction. Since both studies also rely on self-reported metrics, the question formulation from these studies can be used in the method instrument too.

3.1 Exploratory study for measuring anticipated satisfaction

We conducted our first study of anticipated satisfaction as a live experiment at the RefsQ 2012 conference. The study had exploratory character and was our first test for measuring anticipated satisfaction An article describing the study was also published in the post-proceedings of the RefsQ 2012 conference [157].

3.1.1 Research questions

While we are not aware of approaches for measuring anticipated satisfaction, there is a large body of literature on the measurement of actual satisfaction. A popular and simple approach for that is to measure satisfaction-related concepts with a questionnaire, which participants answer after using the system (post-exposure). The obvious way to translate this to an anticipated satisfaction measurement is to measure the same satisfaction-related metrics before the participants have used the system.

The measurement process in this approach consists of three steps:

1. The participants read descriptions of the system's features
2. The participants read questions about the satisfaction-related concepts and answer them, based on the information from the feature descriptions
3. The requirements engineers [1] use the participants' answers to calculate a forecast for the users' actual satisfaction.

These steps provide the research questions for our study.

In the first step, the participants have to understand the feature descriptions. If their understanding is insufficient, they will have no basis for their answers to the questions in the measurement instrument. Our first research question is

- RQ 1: How well do the participants understand feature descriptions?

[1] This is the term we use for the people conducting the measurement project. For more details on the usage, see section 5.1.1.1.

When designing the questionnaire, the requirements engineers have to choose a format in which to represent the feature descriptions. This gives them the opportunity to choose the most understandable format, thus improving the results. The second research question for the study is

- RQ 2: Is there a difference in the understanding of different formats used for the representation of feature descriptions?

The second step requires that the participants answer the questions. This is not a trivial assumption. In our earlier work, we have encountered users who initially agree to participate in a study, but then refuse to give answers. They reported reasons such as the questions being too difficult, or the questions appearing too intrusive, or even because they do not understand the purpose of the study and conclude that it is not worth their time. This leads to the next research question

- RQ 3: Is it feasible to measure satisfaction-related concepts before participants have used the system?

In the last step, anticipated satisfaction is used as a forecast for actual satisfaction. The practical value of this forecast depends on how well the anticipated satisfaction approximates the actual satisfaction. The last research question is

- RQ 4: How well are pre-exposure measurements of satisfaction-related concepts suited as satisfaction measurements?

These research questions cover the major failure reasons of the new approach. If all three steps of the measurement process can be completed successfully, then we assume that the approach is suitable as the basis for a new method.

3.1.2 Materials and methods

Some of our research questions must be answered with pre-exposure measures, such as the measuring of satisfaction-related concepts for RQ 3, while RQ 4 requires post-exposure measures. As our study format was a live experiment where the participants were present for a single session, we had to condense these measures into one questionnaire with two parts and emulate system use for the exposure event.

For the system, we used a fictive software product for managing receipts and recording expenses. The choice of a fabricated system ensured that no participant had previous knowledge of it. The task of recording expenses is simple enough that we could cover the full functionality of the system in a small number of features, and we assumed that anybody who travels to a conference is accustomed to the process of expense claiming, including the need to record expenses documented on receipts.

After some demographic questions, the questionnaire showed the feature descriptions. There was a total of 16 features. As one of the study's goals was to compare different formats, we used two versions of the questionnaire. One included the features as user tasks [123], and the other contained user stories [43]. An example user story described a feature with a sentence written from the user's point of view, like

> The user can import a picture of a receipt.

They were printed in small rectangular frames, to simulate the cards on which user stories are traditionally written.

The user tasks used the same features, but using a different layout and sentence format. The questionnaire reproduced in appendix B is the user task version.

The users were instructed to first read the feature descriptions, then fill the first part of the questionnaire without reading the rest (experimenters in the room ensured compliance). Then they watched a video of the system, then were instructed to fill the second, post-exposure part.

The pre-exposure part contained a measure of a satisfaction-related concept. For this first test, we decided to use a concept from a widespread theory which is simple to understand and apply. Our choice was the *importance-performance analysis* [145], a traditional method for satisfaction measurement with only two concepts. We used the *importance* concept from it as an example for a satisfaction-related concept with which we could test the successful formulation of questions and see if the users are capable and willing to answer questions of this type. The question was worded as:

> I think this feature is ... for the way I will work with the system.

very important		slightly important		not important
☐	☐	☐	☐	☐

The next question in that part measured *perceived understanding*. We asked the users how well they understood the feature descriptions. They had no way to judge the correctness of their understanding at this time, so this question measured not how well they really understood the features, but how well they believed to have understood them. Thus we named the concept *perceived understanding*, to distinguish it from *actual understanding*.

> I can envision a way this feature will be implemented.

really well		unsure		not at all
☐	☐	☐	☐	☐

Both questions were repeated once for each feature, with the text of the feature repeated, but no

longer layouted as in the user story or user task description at the beginning.

The exposure event consisted of the users watching a video. We built a convincing surface prototype of the system described in the questionnaire, and used it to create tutorial videos. These videos were a screencast of a tutor using the system and narrating an explanation of what is happening and why. Each feature used in the questionnaire was present in the videos.

The post-exposure part had three more questions. The first measured *actual satisfaction*:

> I like the feature the way it is implemented now
>
> a lot ☐ ☐ somewhat ☐ ☐ not at all ☐

The two others focused on *actual understanding*. One of them was a standard measure of the concept on a scale:

> The feature implementation corresponds to what I had envisioned
>
> very well ☐ ☐ somewhat ☐ ☐ not at all ☐

The other one was a free-text question whose intention was to give us insights into the causes behind bad understanding:

> The implemented feature differs from what I had envisioned in following ways:

To check how well perceived understanding matches actual understanding, we added a post-exposure part to our questionnaire. It had a question designed to measure *actual understanding* on a Likert-scale, similar to the other questions. Additionally, we also included a free-text question inviting participants to describe how the system differs from their initial understanding of it. A third question in this part was a simple measurement of *actual satisfaction*, allowing us to see to what degree the pre-exposure measurements correlate with it.

The questionnaire also contained a short explanation of the study goals, a few sentences about the system's purpose, and demographic questions. It finished with an invitation for the participants to give feedback. The full questionnaire is reproduced in Appendix B.

At the begin of the 90-minutes session, an experimenter explained the study to the participants, handed out the paper-based questionnaires and they were asked to fill the pre-exposure part. After that part of the questionnaire was filled out, the experimenter presented the video tutorials. The participants then filled out the post-exposure part and returned the questionnaires. The answers were entered into a .csv file, and were analysed using R.

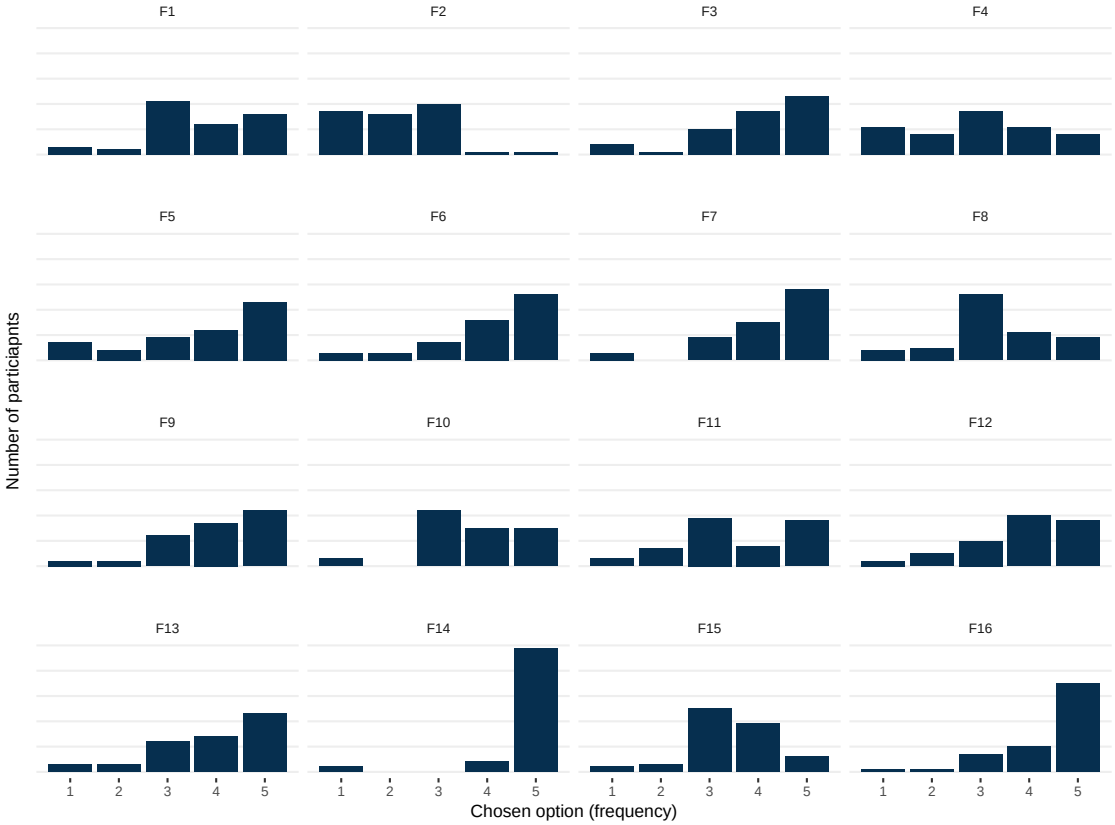


Figure 3.1: Perceived understanding answers for each feature. For the definitions of the 16 features, please see the questionnaire in appendix B

3.1.3 Results

The study was successfully conducted, with 56 participants. We were able to gather information for all three of our goals.

Understanding To answer RQ 1, we used the answers of the *perceived understanding* question and *actual understanding* questions. The distribution for perceived understanding answers for each feature is shown in figure 3.1. We see some variability between features. For most of them, the answers are predominantly on the positive side, with many features showing a heavy increase with a maximum at best option. This shows that participants felt they understood most features well. Also, we see some variation, with some of the features being much flatter than others, like F7, or having a maximum in the middle, like F2, which can be interpreted as users showing much lower satisfaction with F2, and no agreement on F7. This shows that the question allows for good differentiation – it is usable for identifying outliers and investigating how they differ from the rest, if needed.

The *actual understanding* showed mostly two patterns, as depicted in figure 3.2. Some features

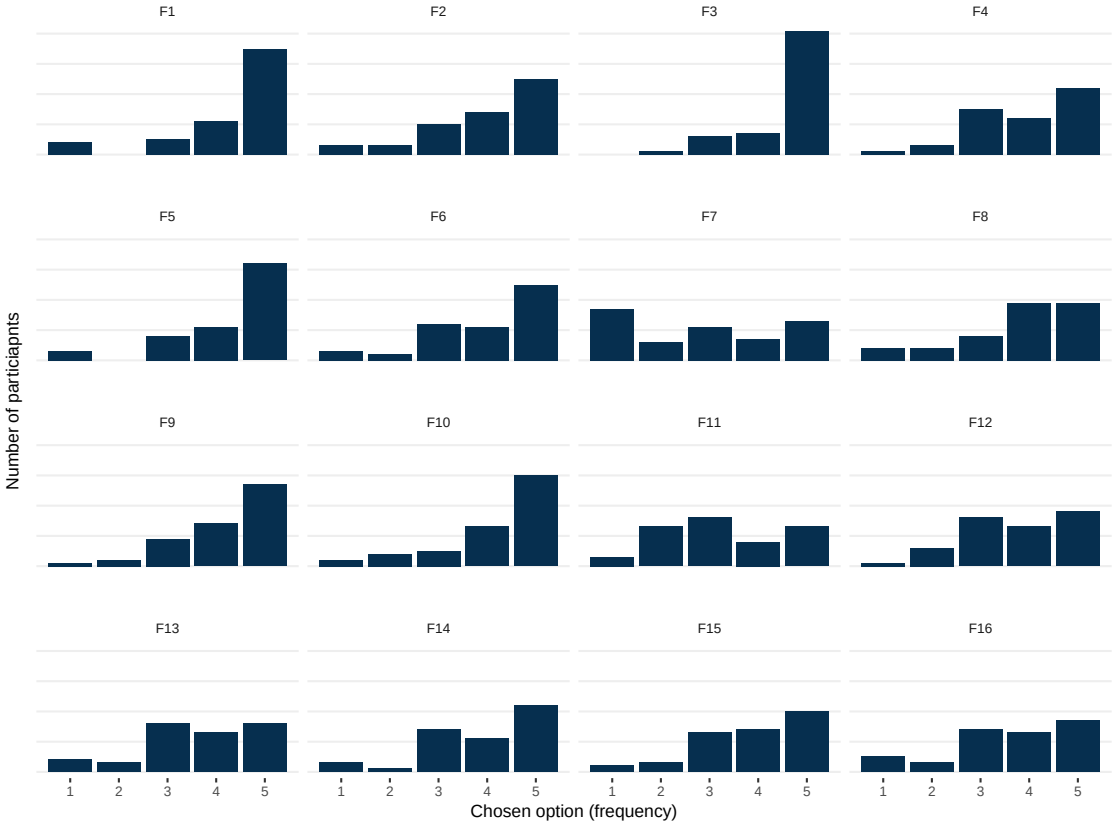


Figure 3.2: Actual understanding answers for each feature

seemed to have been very clear, with a very large number of participants selecting the best option. Others had a very flat distribution, spreading over all five options and with only a slight slope growing towards the better options, which indicates lower understanding. We found a strong correlation between *actual understanding* and *actual satisfaction* ($\rho = 0.5$). There is frequently some difference between actual and perceived understanding (only 32% of the answers show no difference for matched pairs), which confirms that insufficient understanding will be a significant error source in the measurement of anticipated satisfaction. The free-text descriptions of the difference between imagined and real features was left mostly unfilled, so we could not gain any further information from it.

Different formats of feature descriptions To answer RQ 2, we conducted t-tests on the two variables *importance* and *actual satisfaction*, as well as the error in understanding, which we calculated as the numerical difference between actual and perceived understanding, resulting in two tests for each feature. We could only reject the null hypothesis of equality in two of the 32 tests. This is an indication that the difference between user stories and user tasks has no effect on the measurement of anticipated satisfaction. This is convenient for our method, since it it is an indication that a development team which wishes to use it does not have to "translate" their existing requirements into a method-specific format.

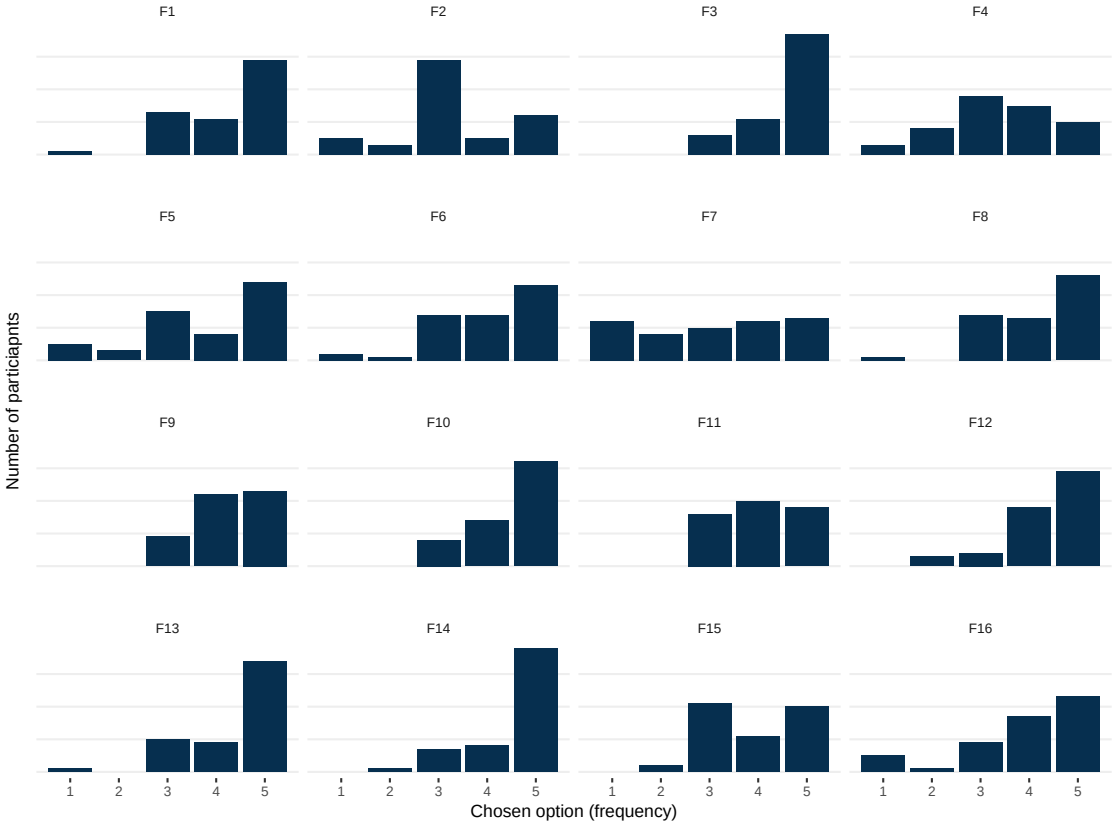


Figure 3.3: Importance answers for each feature

Feasibility RQ 3 asks if users are likely to answer questions on satisfaction-related concepts before they have been exposed to the system.

In this study, we had a 100% response rate. That high rate is certainly due to the setting – we had a captive audience, who had chosen to attend that conference session (although they did not know the details of the study before making that choice). Only one person chose to leave the questions unanswered and to write that the "questionnaire is fundamentally flawed". The remaining 55 participants completed the questionnaires, and while they might have found some questions challenging, there was no indication that they regretted their participation or found it too burdensome, and we also had a good response rate to the optional invitations for freetext suggestions in the end. From that we concluded that participants are willing to cooperate when asked to fill this type of questionnaire.

Pre-exposure measurement of satisfaction-related concepts The last research question was about pre-exposure measures of satisfaction-related concepts. In this study, they were represented by a question about *importance*. The answer frequencies are shown in figure 3.3, broken up by feature. The variable behaves as expected from a satisfaction metric [153] – the most commonly observed distribution is unimodal, and the mode is shifted towards the positive

end of the scale. We also observed differences between the distributions of each feature. For example, Feature 3 was very popular, with 66% of the answers choosing the highest option, and nobody using the negative end of the scale. Feature 11 showed a different pattern, where again the negative end of the scale was not used, but the mode was in the second-highest option, and the used options had almost the same number of answers each. Feature F7 was unpopular, with all options being chosen almost uniformly. This variety of answer patterns indicates that the answers are valid, and can be used to discriminate between features.

We also calculated the correlation of the *importance* variable with the *actual satisfaction* for each feature. The mean correlation coefficient (Spearman) was 0.15 and the standard deviation was 0.12. This correlation is on the low side of the range we see with post-exposure satisfaction metrics, as evidenced by the data from our literature study which we discuss in chapter 4, especially the distribution of all correlation coefficients we found (its quartiles are given in table 4.9). This suggests that it would be useful to search for other metrics which have better correlation with the goal variable.

3.2 Constructing an idealized model of anticipated satisfaction

We used the knowledge derived from our first study to create an idealized model [2] of user satisfaction. We use this model in chapter 5 to derive the questions for the measuring instrument used in MUSA.

A graphical representation of the model is shown in figure 3.4. The rectangles denote concepts, and the arrows are their relationships.

The goal variable in this model is *satisfaction*. Ultimately, the development team is interested in the *actual satisfaction*, but the *anticipated satisfaction* has to serve as a best approximation. The difference between them is the prediction error. We are trying to minimize this error, but it cannot be reduced to zero, since there are concepts which skew the satisfaction measurement before exposure to the system.

The first of those concepts is *perceived understanding*. This is the user's understanding of the feature based on description only. In our theory, it depends on two other concepts, *feature clarity* and *domain knowledge* of the user. These two concepts are important for the theoretical understanding of anticipated satisfaction, but not directly used in the MUSA method. We assume that any available feature description will be already described as clearly as possible, since this is necessary for their primary role in the development project. Thus, a suggestion in MUSA to make features as clear as possible would be pointless - first, actionable advice on how to create and measure clarity would be out of scope of this dissertation, and second, a software

[2] See section 2.3 for a definition of *idealized model* and related terms.

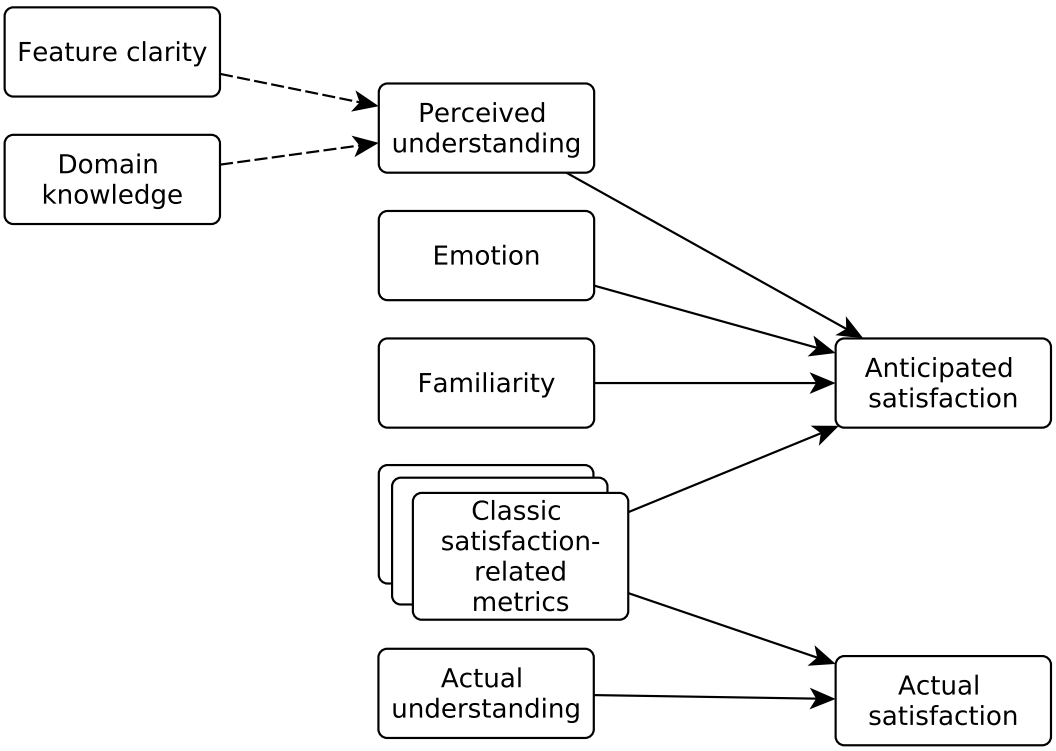


Figure 3.4: A model of anticipated satisfaction

development team which has not written a clear specification for its own use is unlikely to do it for a satisfaction-measuring subproject.

We have anecdotal evidence that the user's domain knowledge, especially knowledge of the process being automated by the system, also plays a role in understanding requirements. However, a development team cannot influence the knowledge level of their user population, and the participants in a satisfaction study should be representative of the total population. Therefore, we have not formally confirmed this connection, and we do not use it in our method guidelines. It is only included in the model to provide a better understanding of the underlying theory.

The next two concepts are *emotion* and *familiarity*. They are empirically derived, and easy to explain theoretically. Satisfaction is itself an emotion, so it requires a person to feel something about the system. If the participants have only neutral feelings, or even feel an absence of any emotion, they cannot report any satisfaction-related emotions. User feedback suggests that these indifferent users are likely to check the neutral option on a scale. This can be useful for some kinds of analysis (e.g. many answers which center around neutral options can suggest the need to proactively create user engagement), but skews the results for others, like ranking the features by user preference.

Familiarity is an expression of the mere exposure effect. This effect has been known to occur in

different contexts, and states in summary that people report a preference for things which they know well over things which are new to them [203]. Especially when their idea of a feature is only based on a brief description, they do not have salient experiences on which to base a firm opinion of the feature, so factors such as the familiarity are likely to have a more prominent role than in a post-exposure measurement.

Our model also contains a multi-node labeled "classic satisfaction-related concepts". We expect that the factors which influence actual satisfaction also influence anticipated satisfaction. Thus, our model assumes that all concepts related to actual satisfaction are also related to anticipated satisfaction. While we cannot prove this for all possible concepts, this assumption seems reasonable, and it holds for the concepts we measured in our own studies.

The reason that we represent these concepts with one multi-node as opposed to listing each of them separately is twofold. First, they all play a similar role within the model, and this representation helps highlight what is different in anticipated satisfaction as opposed to actual satisfaction. Second, there is a large variety of instruments used for measuring actual satisfaction, covering a broad range of concepts, and there is no existing short list which can be included here. In the next chapter, we present a systematic literature review which lists all satisfaction-related concepts which we subsume under the term *classic satisfaction-related concepts* here. This allows the software engineers using our method to choose the concepts which are most relevant for the context of their own project.

The last concept in the model is *actual understanding*. It is measured post-exposure, and provides information on how much the user's imagination differed from reality. It also influences actual satisfaction, for the same reasons that the perceived understanding influences anticipated satisfaction.

The model shows that many of the concepts which influence anticipated satisfaction are only relevant during pre-exposure and lose their effect post-exposure. Thus their influence is contributing to the prediction error, which is the discrepancy between anticipated and actual satisfaction. This includes *perceived understanding*, *emotion* and *familiarity*. This property is important in the context of a measurement method, as it can be used to limit some of the prediction error. The classic satisfaction-related concepts do not fall under this category, as they influence both anticipated and actual satisfaction.

The completed model offers a structured way of understanding anticipated satisfaction. Its focus lies mainly on the ways in which it differs from actual satisfaction, since we have to take these differences into account when deriving our method from classic satisfaction-measuring methods.

3.3 Empirical test of the model of anticipated satisfaction

To validate the model described in the previous section, we conducted a second empirical study with students at Heidelberg University, which was later published in [158]. For this study, we expanded our instrument and translated it to German, then administered it in a way similar to the study described in section 3.1.2.

- RQ 5: Are the relationships predicted by the model of anticipated satisfaction empirically observable?

From this question, we derived six hypotheses. Each of them corresponds to a relationship in the idealized model of anticipated satisfaction.

Hypothesis 1 Perceived understanding of a feature is related to anticipated satisfaction with that feature.

Hypothesis 2 Familiarity with a feature is related to anticipated satisfaction with that feature.

Hypothesis 3 Strong emotions about a system lead to better differentiated anticipated satisfaction with the features of that system.

Hypothesis 4 Anticipated usefulness is related to anticipated satisfaction.

Hypothesis 5 Actual understanding is related to actual satisfaction.

Hypothesis 6 Different feature formats do not have an influence on anticipated satisfaction.

3.3.1 Materials and methods

We announced the study at Heidelberg University, and all participants were undergraduate students (N=112). The process was similar to the live experiment - in a single session, the participants read a list of features, filled the pre-exposure part of a questionnaire, watched a demonstration of the system, then filled the post-exposure part. The system was the same as used in the live experiment, a prototype of a receipt manager usable for expense tracking. The students received 20 euro for their participation.

The pre-exposure part measured three variables – *perceived understanding*, *anticipated satisfaction* and *usefulness*. *Perceived understanding* was analogue to the question in the first study. *Usefulness* represented the *classic satisfaction-related concepts* in the idealized model and was measured with the question (translated from German)

> I can imagine ... why this feature is needed and how I would use it.

clearly		vaguely		not at all
☐	☐	☐	☐	☐

Anticipated satisfaction was measured with the question

> When the feature has been implemented, I will have following feeling

I will like it		I will be indifferent		I will not like it
☐	☐	☐	☐	☐

Similarly to the last study, the students did not use the prototype, but were shown a video. It was structured as a tutorial of the receipt management system, showing in a screencast how to go through the process of scanning a receipt, extracting its data and saving it in an expense tracking sheet, accompanied by a narration. It had to be created anew for this study, because this narration was in German. The script for the individual steps and their order was the same as in the first study.

The post-exposure questionnaire measured *actual understanding*, *actual satisfaction* and *familiarity*, once for each feature. The variable *emotion* was measured once for the whole system, as we did not expect the participants to have fine-grained feelings which differ between features. There was again one question per concept, directly naming the concept in the question formulation. The questionnaire ended with a free-text feedback field.

The two new questions were *familiarity* and *emotion*. The wording for *familiarity* was

> I have used a similar feature in a different software system before

Yes, very similar		Yes, somewhat similar		No
☐	☐	☐	☐	☐

For *emotion*, we asked

> I found the notion of using this system

Awesome		Boring		Nonsensical
☐	☐	☐	☐	☐

We also conducted a comparison between three different feature formats – the *user stories* and *user tasks* already used in the live experiment, as well as features described with sentence templates as suggested by Rupp [161]. An example for the third format is

> The system should be capable of recognizing the text in the scanned image.

The questionnaires were printed on paper and given to participants to fill offline, randomly assigning participants to one of the three task formats. The responses were transcribed into a .csv file and analyzed with R.

3.3.2 Results

As this was a confirmatory study, the analysis method differed from the analysis of the live experiment, even though the questionnaires were similar. In this study, we used hypothesis testing to evaluate the strength of evidence for the relationships in the model presented in the previous section.

When we used a correlation coefficient as evidence, we tested the null hypothesis that there is no correlation ($\rho = 0$). The significance level was $\alpha = 0.05$ for all tests. The p-values we report correspond to this test.

Hypothesis 1 The mean correlation of *perceived understanding* and *anticipated satisfaction* over the 16 features was $\rho = 0.5$, with a standard deviation of $0.16, p < 0.001$. This is a strong correlation, and we see it as an indication that the hypothesis is correct.

Hypothesis 2 For *familiarity* and *anticipated satisfaction*, the correlation was $\rho = 0.16$ and its standard deviation was 0.07. This value is significantly different from zero ($p = 0.048$), which we see as evidence that familiarity has an influence on anticipated satisfaction, even though it is weaker than that of perceived understanding.

Hypothesis 3 This hypothesis states that participants who feel a strong emotion about the system will give better differentiated answers than those who feel little emotion. To investigate that, we calculated the standard deviation of each participant's anticipated satisfaction across the 16 features. For participants who gave the highest possible score on *emotion*, the median standard deviation was 1.01. On the other extreme, participants who had the strongest negative emotions had a median standard deviation of 1.02. The participants who chose the central option on the emotion scale (indifferent to the system) had a median standard deviation of 0.96, which means their answers had less diversity than those of participants with a strong emotion in either direction. The difference is small, but it is in the direction which our theory predicts. We see this as evidence that our hypothesis might be correct, although it is not very strong and we cannot consider it confirmed.

Hypothesis 4 The *usefulness* we measured in pre-exposure was well correlated with anticipated satisfaction, $\rho = 0.61$, with a standard deviation of 0.14, $p < 0.001$. This is a strong

correlation, and we regard the hypothesis as confirmed.

Hypothesis 5 *Actual understanding* exhibited a good correlation to *actual satisfaction*, $\rho = 0.38$, standard deviation 0.2, $p < 0.001$. We see this as evidence that the hypothesis is correct.

Hypothesis 6 As in the earlier study, we conducted t-tests for each feature, with the null hypothesis that the average *predicted satisfaction* does not differ between different feature formats. From the 48 resulting t-tests, only one could reject the null hypothesis. As we did not correct for the familywise error rate, this single result is likely a false positive. We conclude that there is no difference in the anticipated satisfaction between participants who read features in different formats.

In conclusion, we could find good evidence for four of the six hypotheses. The evidence in the other two (hypothesis 2 and 3) is aligned with what the hypothesis predicts, but it is too weak to be seen as a confirmation.

3.4 Discussion and conclusions

With our second study, we were able to find evidence confirming our idealized model of anticipated satisfaction. We followed standard academic practice by starting with an exploratory study, creating a model, then confirming it with a second study. Nevertheless, our work has its limitations, which we discuss in the next subsection. We conclude this chapter with a discussion of the role this model has in creating the MUSA method.

3.4.1 Limitations

The work represented in this chapter represents the early stages of creating a theory of anticipated satisfaction. We did not have the resources or need to create a full theory for our method, so our findings have a preliminary character. There are several threats to the validity of our work. We discuss here threats referring to all conclusions made in this chapter, including results from both studies and the idealized model itself.

Conclusion validity The main threat to conclusion validity is the problem of multiple testing. In the second (confirmatory) study, we used a single dataset to test 6 hypotheses. Thus the probability that the results we observed are due to chance is higher than the chosen significance level implies.

Furthermore, the data from that second study showed only weak correlations for two of the relationships in our model. More research is needed to establish the true nature of these relationships.

Internal validity The quality of the instruments we use in both studies may have compromised the validity of our findings. We were not able to use existing, validated instruments. Instead, we created our own. To mitigate the risk of ambiguous wording and other quality problems, we piloted them with colleagues and edited them before using them in the studies.

Construct validity Ideally, a psychometric measure for a given concept would measure multiple dimensions for it, and use an instrument which was created and refined through multiple studies. Creating such instruments for our concepts would have been far outside the scope of this dissertation, so we opted to measure each concept with a single variable with the same name as the concept. This approach is frequently used in satisfaction measurement.

External validity The context of our studies was not entirely realistic. The participants were not sampled from a population which has expressed previous interest in the system. Also, they did not complete a task with the system, so their experience is different from that of typical users.

These limitations mean that we cannot yet regard our model and the underlying theory as fully established. Nevertheless, we find that the strength of the evidence is sufficient for deriving a measurement method, which can be employed in a practical setting. Further research is needed to fully test the model and, if necessary, refine it.

3.4.2 Conclusions

Our first goal in this chapter was to gather first experiences in measuring anticipated satisfaction, the second one was to create a model of anticipated satisfaction. We conducted two empirical studies for this. The first had exploratory character, measuring a satisfaction-related concept and several other variables designed to indicate whether the measurement was successful. We used the information from this study to create a proposed idealized model of anticipated satisfaction. The second study tested the relationships in that model. There was strong evidence for the four of the six relationships in the model, and weak evidence for the correctness of the remaining two. This model fulfils the second goal of the chapter.

Conducting the studies led to several important insights about measuring anticipated satisfaction. With them, we have achieved the first goal of the chapter. These insights inform the development of a measuring method of anticipated satisfaction in later chapters. We list them here in summarized form.

One central finding of our research is that anticipated satisfaction is influenced by two groups of concepts, the classic satisfaction concepts (those which also influence actual satisfaction) and the concepts responsible for prediction discrepancy. If a practitioner were to only measure the satisfaction concepts, he or she would have no basis of knowing how different the results are from actual satisfaction. Therefore we recommend to measure both types, and to use the data from the discrepancy-causing concepts to judge the reliability of the satisfaction concepts. If enough data points are available, it is even feasible to filter the data and only use satisfaction data from users with a favorable discrepancy profile. For example, answers from users who indicate low understanding of the system description can be excluded from evaluation.

This implies that a questionnaire should contain both satisfaction questions and discrepancy questions. The three discrepancy questions to include are those about *understanding*, *familiarity* and *emotional attachment* to the system. While our model differentiates between *perceived understanding* and *actual understanding*, it is not practicable to measure the actual understanding. Suggestions for the use of discrepancy data in satisfaction evaluation are elaborated in chapter 5.

Both the satisfaction questions and the discrepancy questions in our study were well-received and resulted in useful measurements. Therefore we can recommend using the same formulation when applying our method.

We also included open ended questions successfully. They are much more work intensive than closed questions, but also provide valuable information. We believe that the best use for our method is to improve the requirements in places where users are dissatisfied. Gathering their improvement ideas may lead to creative solutions, which justifies the increased effort.

The above points all refer to the first research question in this chapter, which is about choosing and formulating the questions in a measurement instrument. The second research question refers to the representation of the requirements, which we also investigated in both studies presented here.

Our method places the system requirements in the questionnaire itself. The requirements should describe features of the system, as we have not tested other artefacts of the requirements specification (e.g. personas). Our research found no difference between users receiving different requirement formats, so we suggest that practitioners use the format already present in their project. This makes the method compatible with a wide range of development process types, lowering the barrier to adoption. We also recommend adding a brief description of the system's purpose and overall use at the beginning, to give the users the necessary background to understand the detailed requirements later.

The method we recommend borrows much of the ideas we applied in this chapter. However, the questionnaires we used are not intended for direct use by practitioners. First, they include both a part which measures anticipated satisfaction and a part which includes a measure representative of actual satisfaction. Second, they are not complete. We used a single question to represent the classic satisfaction concepts in our model. For a complete measurement, several satisfaction

questions are needed per feature. The next chapter concentrates on that, delivering a list of satisfaction related concepts which can be used to formulate questions for our method.

4 Satisfaction related concepts

In our method for anticipated satisfaction, a questionnaire contains metrics representing both the discrepancy concepts identified in last chapter and a number of satisfaction-specific concepts. The last chapter focused on the discrepancy between anticipated and actual satisfaction, and the concepts explaining it. This chapter investigates the role of the satisfaction related concepts, and their use as metrics. Its goal is to provide a list of concepts which can be used to derive questions for our method.

This list should consist of metrics representing concepts known to be correlated to satisfaction when measured with a questionnaire. To facilitate use, it should be organized in a way which provides a good overview and allows finding of related metrics. Furthermore, as practitioners will be choosing metrics from that list, it would be useful to have an estimate of the strength of relationship between each connection and satisfaction. This results in the two research questions:

- RQ6: Which concepts have been studied in relation to user satisfaction?
- RQ7: How strong is the relationship measured between user satisfaction and related HCI concepts?

We consider each concept that is used for measuring another concept to be a metric [1]. Measuring a satisfaction-related concept with multiple metrics is possible, but falls outside of our scope due to the large volume of such concepts. It is not feasible to investigate multi-measurement for them within a dissertation, and also our questionnaire does not foresee the use of multiple metrics per concept, as this would create questionnaires too long to be used in practice. Thus we only consider a single metric per concept (a simple one-to-one relationship) and call the metric and the measured concept with the same name.

As there is a large body of research investigating satisfaction measurement, our approach for answering these question is a literature study. We chose to conduct a systematic literature review, which provides good evidence quality, captures the important developments in the field, and allows a quantitative metaanalysis. Our review encompasses studies which empirically confirm a model for measuring user satisfaction. The metrics in these models are used to answer the first research question, and for those which have a sufficient number of correlation measurements to satisfaction, we can provide a metaanalytic estimate of the relationship strength.

[1] This use is based on a definition for the term *metric* which is specific to software engineering, as described in chapter 2.4.

We conducted a pilot study before the full literature review. Its major result is a categorization schema for satisfaction metrics, described in section 4.1. It is used to organize the metrics found later in the study. Section 4.2 describes our study protocol. The metrics found in the review are described in section 4.3, while their relationship to satisfaction is the focus of section 4.4. It is followed by a discussion, which explains the relevance of the review for the method developed in later chapters. The concluding section makes recommendations how to use the results of the systematic literature review when developing a new questionnaire.

4.1 ASMA categorization schema

Before initiating a systematic search, we carried out an explorative pilot study without strict search and selection criteria. Its goal was to get acquainted with the available literature, enabling us to design a better systematic study. Roughly 200 articles mentioning satisfaction were used for the pilot study.

If a primary source in this study had a data model of user satisfaction, the variables in that model were extracted and treated as concepts of an idealized model, as described in section 2.3. It soon became apparent that they are highly heterogeneous, with different authors investigating similar concepts or even the same metric under different names. We estimated that the systematic study will yield dozens of distinct metrics. In order to structure them, we decided to create a categorization schema.

4.1.1 Schema derivation

For the categorization schema, we used a combination of a data driven and a theory driven approach. The data driven approach used the metrics from the pilot study as a source. The theory driven approach used the basic theory underlying human-computer interaction, as described for example in [54].

4.1.1.1 Data driven approach

During extraction, we noticed that metrics are sometimes similar, and started noting a category candidate for each. The category candidates were terms we felt described the metric well and were likely to apply to other metrics.

As an example, the metric *use*, which describes how often the user uses the system, was mapped to the category candidate *user behavior*. The metric *fun* was categorized as *user affect*, as fun is part of the affective state of a person. *Price* was seen as belonging to *finance*.

After the data extraction was finished, the resulting category candidates were compared. Inconsistencies were cleaned and better fitting names chosen. The resulting categories had several advantages. They were disjoint, the granularity seemed right, as there were neither too many categories nor too large differences between the metrics in the same category. At the same time, they also had drawbacks. The criteria for assigning a metric to a category were fuzzy, and the categories bore little relation to each other. While some of them obviously had something in similar, e.g. *user behavior* and *user affect* both pertaining to the user, most of them came from varying backgrounds. For example, *user affect* is a psychological concept, while *finance* is a topic in economics. As a result, the set of the category candidates we derived from the data was difficult to work with, as it lacked a coherent structure.

4.1.1.2 Theory driven approach

In the next step, we decided to start with an established theory and check whether its tenets can be used to categorize the metrics from our dataset. The subjects of our analysis are scientific studies involving a satisfaction measurement.

A user satisfaction measurement is based on a situation of human-computer interaction. The components of the system of interest are real world entities. Figure 4.1 contains an informal representation of the system of interest for a user satisfaction measurement. The main actors are the *user* and the *software system*. The user performs a *task*, and interacts with the software system for that purpose. The software system processes the *information* needed to complete the task. The user feels some degree of *satisfaction*. This is embedded in a social and organisational context, which places constraints on the actors and their interactions. For example, a user who performs a task for his or her employer may be required to use a software system provided by the employer.

As an initial step in our theory driven approach, we compared the category candidates from the data driven approach with the entities which comprise the system of interest, and we discovered that each of these category candidates corresponds to one of these entities. There were no candidates which we could assign to the *task*, and *satisfaction* is an attribute of the user, leaving the entities *user*, *system* (short for software system), *information* and *context* (short for organisational and social context).

Having only four categories would have resulted in too coarse granularity for our purposes. We decided to subdivide the category candidates by further principles beyond their describing an entity from the system of interest.

A standard principle in computer science is that an entity can be described by its properties and its actions. This separation is derived from metaphysics and is reflected in many modelling and programming paradigms, such as object oriented programming and entity relationship modelling for databases [170]. We noticed that the majority of the data-derived category candidates can

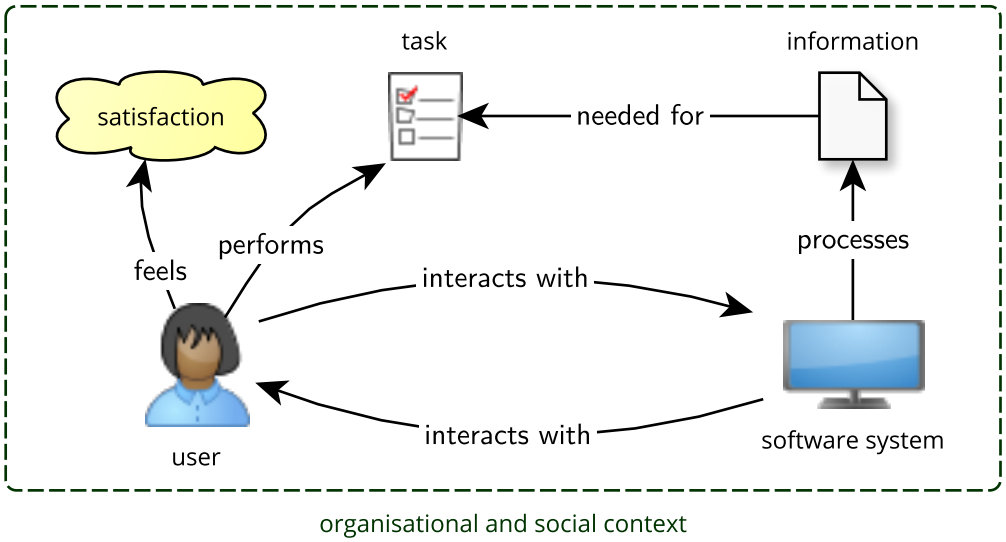


Figure 4.1: An informal representation of the system of interest in measuring user satisfaction with its components

also be sorted as pertaining to either actions or properties. Some of these properties would be constant over multiple measurements, for example a user's *experience with technology* should not change between two separate measurement sessions, while the user's *emotion* is likely to change. We considered these to be a *stable property* and a *mutable property* respectively. We chose the term *activity* to denote the collective sum of an entity's actions.

Beside metrics describing an entity's activity and its properties, there were ones which described a judgement of the entity. We did not consider them to be a property of the system, since they do not exist without an interaction between the user and the system, and can differ between different users. Instead, we considered these to form a fourth type of category, and classified them as *appraisal*.

4.1.2 Consolidation and final form of the ASMA categorization schema

We see activity, stable properties, mutable properties and appraisal as different *dimensions* which contain HCI concepts measurable by a variable. They are orthogonal to the entity to which the metric belongs. The combination of an entity and dimension determines the category of a metric. We termed this categorization the *ASMA categorization schema*, short for Actions, Stable properties, Mutable properties and Appraisal.

While a category is determined by both an entity and a dimension, not every combinatorial pairing of an entity with a dimension produces a valid category. Several pairings make no sense

semantically, and thus are not considered ASMA categories. This section lists the valid categories sorted by entity.

User The user entity represents a human being, and has activity, stable properties and mutable properties. Judging the user would be questionable from an ethical viewpoint, and in practice, we found no *user* metrics which would fall under the *appraisal* dimension. Thus, the three valid categories for this entity are *user activity*, *user stable properties* and *user mutable properties*. The central concept in our research, *user satisfaction*, belongs to the category *user mutable properties*.

System This entity represents the software system with which the user interacts. It has activity and properties, and can be appraised. However, while it is theoretically possible that a software system has mutable properties, our pilot study found no metrics which would fall into that category. So we consider it to not be part of ASMA. Therefore, the ASMA categories for metrics describing the software system are *system activity*, *system stable properties* and *system appraisal*.

Information Information is a passive entity being processed by the software system. It has no activity of its own. Its ASMA categories are *information stable properties*, *information mutable properties* and *information appraisal*.

Context While the three other entities in ASMA have a concrete pendant in the real world, the context is much more general. It stands for anything which is not an inherent part of the system of interest, but nevertheless has an influence on the user satisfaction measurement. It does not participate in the interaction between the user and the software system, so it has no activity. We did not encounter metrics representing an appraisal of the context. So the two categories relevant to ASMA are *context stable properties* and *context mutable properties*.

The eleven resulting ASMA categories are depicted in figure 4.2. Each black-bordered rectangle represents an ASMA category, sorted vertically by entity and horizontally by dimension. An empty intersection of entity and dimension means that no category exists for this pairing.

The ASMA schema derived from the exploratory study was used for organizing the data collected in the full literature review.

4.2 Research method

For our literature study, we followed the guidelines suggested by Kitchenham. [112]. We created a review protocol before the search, including a definition of the search terms and sources to

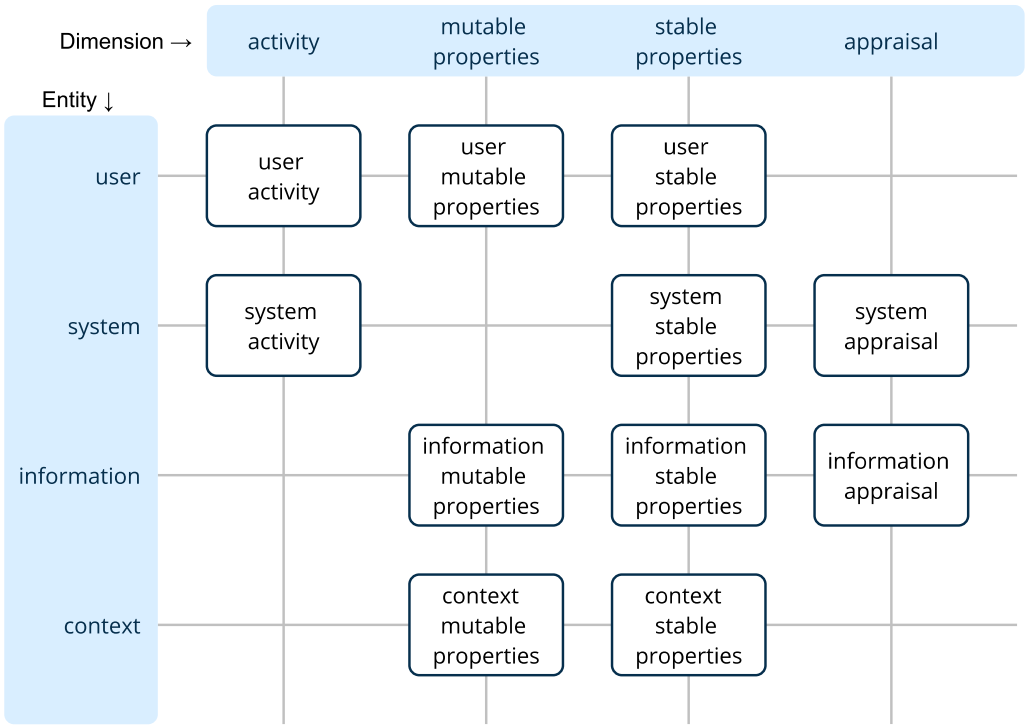


Figure 4.2: The categories in the ASMA schema, sorted by entity and dimension

search for primary studies. The protocol also defined inclusion and exclusion criteria. We documented the search process in detail, creating a full list of the excluded studies with an exclusion reason.

The guidelines also require that quantitative and qualitative studies should be analyzed separately. As the inclusion criteria required a specific type of quantitative research, there were no included qualitative studies to analyze.

One exception to the guidelines was that the primary study selection was done by one person only. However, the main inclusion criterion was not subjective, and all studies which met that criterion were included.

All data was stored in flat .csv files to ensure compatibility and portability. Careful manual linking allowed us to maintain data quality despite this unusual format. After the data was extracted and prepared, we applied a statistical analysis to answer the research questions stated at the beginning of the chapter. For this analysis, we created a custom tool in R.

4.2.1 Search strategy

There is a vast literature body concerning user satisfaction. To answer RQ 6, we needed sources which provide an idealized model of user satisfaction. For RQ 7, the sources also had to provide a data model measuring the relationship strength. From these two requirements, the second one had more practical effects on our search. As a source's idealized model can be derived from its data model, we were able to include sources which do not report an explicit idealized model, and to still use them for RQ 6.

We searched for primary studies using two types of data source, publication databases and manual search of peer-reviewed periodicals. For the database search, we used our insights from the pilot study to design a suitable query. For the manual search, we chose a list of relevant periodicals and read the table of contents of all issues to choose inclusion candidates. We then used a list of inclusion and exclusion criteria to select the final set of primary studies for the literature review.

4.2.1.1 Database search

The research questions required us to find publications which describe the relationship between satisfaction and metrics of other HCI concepts. Explorative searches revealed that most of the articles on satisfaction describe the authors performing a satisfaction measurement, but without investigating the relationship satisfaction has to other metrics. There were many thousands of articles returned for simple queries like *user AND satisfaction AND measurement*. The articles which contained the information needed for our literature study were the ones in which the authors explicitly stated a new idealized model of satisfaction and evaluated it empirically, or created a new instrument for measuring satisfaction through related concepts. Thus, we constructed a search query to find papers which are likely to describe an instrument for measuring satisfaction, or the idealized model underpinning it. We found several terms to distinguish these papers: *factor* (as in "factors which influence satisfaction", *model* (as in "a model of user satisfaction"), *questionnaire*, *metric* and *instrument* (as in "an instrument for measuring user satisfaction"). We restricted these terms to the abstract, as they appeared too often in the body of articles irrelevant to the search. Searching for them in both title and abstract did not improve the search.

A second problem arose due to the different meanings of the concept of *satisfaction*. Searching for it returned more false positives than suitable articles. Many of them were from the realm of theoretical computer science and were concerned with satisfiability problems. Others were about job satisfaction or pupils' satisfaction with an e-learning course. Restricting it to the exact phrase *user satisfaction* would have created too many false negatives, as authors who write about user satisfaction do not consistently use the descriptor. The solution we used was to require that the title contains the word *satisfaction*, while the word *user* is present in the abstract. We found that in user satisfaction papers, users are usually mentioned in the abstract, even when the exact phrase *user satisfaction* is not used.

No restriction by year was used, as satisfaction models are not coupled to technology and the older verified models are still relevant today.

Thus, the final search query was:

((Abstract:user) AND (Title:satisfaction)) AND (Abstract:factor OR Abstract:model OR Abstract:questionnaire OR Abstract:metric OR Abstract: Instrument)

We conducted the database search in two databases specializing on software and computing, *ACM Digital Library* [3] and *IEEE Xplore* [94]. Due to the popularity of the satisfaction concept and the lack of consistent word usage, searches in multidisciplinary databases returned mostly false positives and were not used for the final selection.

The search yielded 241 results from the ACM Digital library and 398 results from IEEE Xplore. After removal of duplicates present in both databases, the total amount of study candidates found in this search was 563.

4.2.1.2 Manual search

Beside a database search, we also performed a manual search, which covers literature not present in the databases we used. User satisfaction measurement is considered a type of usability measurement, thus we needed to cover appropriate publications. Usability researcher Jeff Sauro has listed in a blog post 17 major peer-reviewed periodicals which publish research on usability measurements [165]. We adopted this list for our manual search.

We reviewed the table of contents of all issues of each of the 17 periodicals, looking for articles which might contain a model of user satisfaction. If a title looked promising, we also read the abstract of the article. If there were no signs that the article focuses on something else, we included it in our list of search results.

Table 4.1 lists the number of study candidates found for each of the thirteen journals and four conference proceedings we searched, after excluding study candidates already found in the database search. We found a total of 494 study candidates in the manual search.

4.2.2 Inclusion criteria

The two searches yielded a total of 1057 study candidates. Based on the exploratory study, we created a list of inclusion criteria to select those suitable for the review.

For RQ 6, we needed the names of HCI concepts which have been linked to user satisfaction. RQ 7 had to be answered based on a description of relationship strength. To answer these questions,

Publication	Type	Studies chosen
Applied Ergonomics	journal	17
Behaviour and Information Technology	journal	65
CHI	conference	45
Communications of the ACM	journal	10
Computers in Human Behavior	journal	91
Ergonomics in Design	journal	2
HCII	conference	15
Human-Computer Interaction	journal	10
Human Factors	journal	5
Human factors and Ergonomics	conference	82
INTERACT	conference	21
Interacting with Computers	journal	38
Interactions Magazine	journal	5
International Journal of Human Computer Studies	journal	30
International Journal of Human-Computer Interaction	journal	45
Journal of Usability Studies	journal	5
User Experience Magazine	journal	1
Total		494

Table 4.1: Results of manual search

we needed studies reporting a data model of user satisfaction. The following inclusion criteria ensured that the selected studies contained the information required for our research questions.

Study must be about user satisfaction As explained in the last section, the word *satisfaction* can be used for different concepts. Several of our search results were false positives, focusing on the satisfaction of mathematical constraints, or on people being satisfied with something other than a software system, such as the job satisfaction of software developers. We removed 207 such false positives from the pool of studies.

Satisfaction must be research focus For many articles, the term *satisfaction* appearing in the title or abstract referred indeed to user satisfaction. But upon reading the full text, we realized that the research described in the article does not include the measurement of user satisfaction. An example would be a paper presenting a new system and claiming that the system leads to high user satisfaction, but not presenting evidence for that. We removed 382 articles which did not contain a satisfaction study.

Article must describe an empirical study A data model is based on empirically collected data. Articles which did not have an empirical measurement of user satisfaction – for example metastudies, or suggestions of a new theory or a new idealized model without validation – were removed. 52 articles belonged to this category.

Not a simple system evaluation Some articles were not focused on advancing the scientific knowledge of user satisfaction, but on evaluating a software system by several parameters including user satisfaction. They included an empirical study on user satisfaction, but only reported the final measurement of each variable, without fitting a data model. 61 search results were excluded because of this.

For this criterion, we focused on the reported information and not on the researcher's purpose. Thus, if the researchers had conducted the study for the purpose of evaluating a system, but reported relationship measurements in their article, the study was not excluded.

Must contain a suitable model of user satisfaction For our systematic literature review, we had to extract a data model from each primary study and derive an idealized model from it. We included both studies which described a model explicitly and ones which contained all needed information without representing it as a single compact model. If it was impossible to extract a data model, or that model did not include satisfaction, or it did not include measurements of relationship strength, we did not include the study.

For this criterion, the statistical methods employed in a study played an important role. As a rule, different measurements of relationship strength cannot be compared directly. We had to restrict the types of measurement in our dataset to be able to do comparisons. We chose to include studies reporting either a correlation coefficient, a path coefficient derived from a structural equation model, or a regression coefficient. We excluded studies which used some other, rare form of strength measurement. We also excluded studies doing a factor analysis of satisfaction, as they do not contain a measurement strength of the relationship between the studied concepts and the superordinate concept.

We excluded a total of 164 articles which did not contain a model of user satisfaction meeting the conditions described in this criterion.

Applying all inclusion criteria to the search results left 191 studies to be included. Figrue 4.3 summarizes the proportion of studies which were excluded for not meeting a given criterion.

4.2.3 Exclusion criteria

Some of the studies which met our inclusion criteria could not be used for technical reasons, or were not of sufficient quality. We created a list of exclusion criteria which we used to remove unsuitable studies from our pool.

Vanity press The search results consisted of a name and abstract only, while access to the full text was available through a library subscription in most cases. We noticed that, among

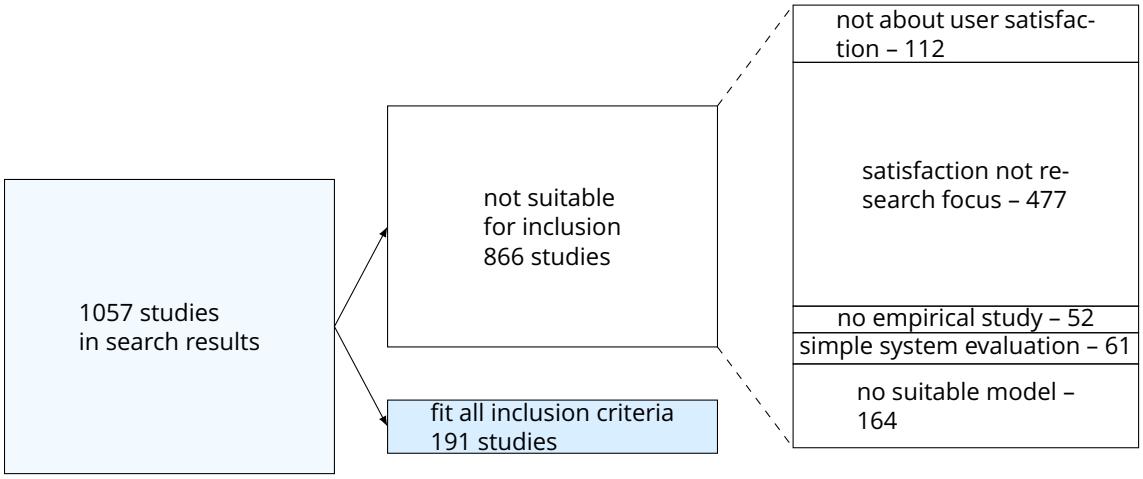


Figure 4.3: Number of articles which fit the inclusion criteria

the publications not accessible through our library, many came from journals belonging to the same few publishers. On researching these publishers, we found for one of them predominantly negative publicity and accusations of being a vanity press and having a very low quality of peer review. We decided to exclude their publications from our study, accounting for 12 study candidates.

Not obtainable We used our own library access, freely available articles, and a library interloan service to obtain the full text of our study candidates. Despite using multiple channels, we could not obtain 8 of our search results. They were all dissertation theses. At 0.8% of all search results, this represents a very low loss of potential sources.

Study already included Sometimes researchers do several evaluations on data from the same study and publish them in separate articles. If we had included relationships found in the same dataset more than once, this would have skewed our results. Therefore we decided to remove articles when we recognized that they describe a study already included in our pool of primary sources. 8 articles were excluded for this reason.

Insufficient data Our search results included a few short articles describing a study which would have met our inclusion criteria. They reported that a data model had been constructed. However, the detail included in the article text was not sufficient, and it was impossible to extract the data model. 32 such articles were removed.

After the sorting, 131 articles were selected for the literature review. The remaining ones were excluded because they fit one of the ten exclusion reasons listed above. Figure 4.4 presents a graphical summary of the distribution of included and excluded articles.

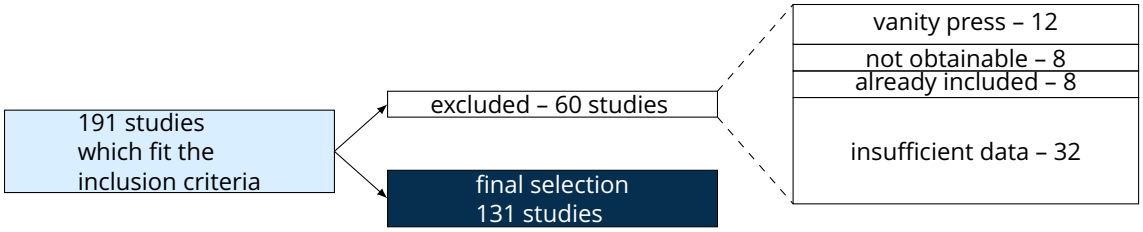


Figure 4.4: Number of articles which were removed by exclusion criteria

4.2.4 Evidence quality

A literature review should establish the evidence quality in the primary studies. Software engineering has no strict evidence standards defined for empirical evidence. We used the criteria suggested in Kitchenham's guidelines [112]. From her list, we selected five criteria which are suitable for surveys and which showed interesting variability in the pilot study. Most of them were coded with three levels, corresponding to good, average and poor quality, while the number of participants had a continuous measure. Table 4.2 summarizes how well the studies met the criteria.

Sample representativeness Fully representative, narrowly sampled (participants are drawn from a small subset of the desired population such as employees of one company), partly representative.

Significance reporting Exact p-value, significance level, or none.

Variable definition All, some or none of the variables from the study are defined.

Limitations discussed Yes or no. We coded *yes* for an actual discussion present, regardless of whether it was presented as a separate section or included in another section. There was no middle level for this criterion.

Number of participants The number of participants who returned valid questionnaires.

Criteria	Number of studies		
	Good	Average	Poor
Sample representativeness	61	28	36
Significance reporting	13	87	30
Variable definition	20	73	38
Limitations discussed	90	–	41

Table 4.2: Evidence quality of the primary studies. Each row lists how many studies fell into each level of a given quality criterion.

The overall quality of our primary sources is fair. The situation is especially good for sample

representativeness, where even the poor studies use a population which is at least partly representative for the desired population, and the majority of studies use a highly representative sample. Limitation discussion is also predominantly good, with roughly ⅔ of publications including a discussion. The authors are less diligent in providing definitions, with the majority only giving definitions for some of the variables they use, and a substantial number dispensing with definitions altogether. Significance reporting is the worst of the four coded criteria, with only 13 publications providing exact p values.

We also included the number of participants as a quality criterion. As it is measured on a continuous scale, it cannot be readily compared with the other four. We cannot provide cutoffs for *sufficient* or *insufficient* participant numbers, as they vary depending on the study design and the researchers' goals. A rough indication can be gained by employing power analysis techniques.

With the exception of six large studies [26, 12, 13, 91, 144, 206], the sample size of the primary studies is below 1000 participants and the median study has 272 participants. At these sizes, many of the primary studies cannot distinguish weak effects from zero [2]. This means that, while we can aggregate the findings of the primary studies, our results are associated with a somewhat high measurement error.

4.2.5 Data extraction

From each primary study, we extracted data on three levels: study, variables and relationships. The study level records data concerning the study as a whole, such as the title and author. On the variable level, we noted each variable from the study's data model. For each relationship between two variables, we created a record on the relationship level, indicating the strength as measured in the study. We stored each as flat records in a .csv file, using a short unique string to identify each publication and the concepts and relationships belonging to the study in it. The following subsections describe the structure of each data level.

Running example In order to illustrate our approach, we chose one of the articles in our review as a running example. For each data extraction and data transformation step, we show the information based on this one study in the text. The complete dataset as extracted from all studies is not part of the printed version of this thesis, but will be available online.

The chosen study is titled "Perceived fit and satisfaction on web learning performance: IS continuance intention and task-technology fit perspectives" [138], and is in many ways representative for our dataset. It is an article by Wen-Shan Lin published in 2012 in the International Journal

[2] The formula for a correlation sample size is $N = \left(\frac{(Z_\alpha + Z_\beta)}{C}\right)^2 + 3$ where N is the sample size, Z_α the quantile of the normal distribution for the desired Type I error rate, Z_β the quantile of the normal distribution for the desired Type II error rate, and C the correlation coefficient measured [93]. If we accept a Type I error rate of 0.05 and Type II error rate of 0.2, a study would need 68 participants for a correlation coefficient of 0.3 to be significant, 153 for a correlation coefficient of 0.2 and 617 participants for a correlation coefficient of 0.1.

Title	Perceived fit and satisfaction on web learning performance: IS continuance intention and task-technology fit perspectives
Authors	Lin, Wen-Shan
Year published	2012
Found by	manual search
Published in	International Journal of Human-Computer Studies
System type	e-learning
Short description	A SEM combining task-techonogy fit and TAM
Interesting for	Includes both SEM and factor analysis.
Feature level measurement	no

Table 4.3: Study level data example for a sample article [138]. SEM = Structural Equation Model (a statistical method), TAM = Technology Acceptance Model (a theoretical model of user satisfaction)

of Human-Computer Studies, and was found by manual search. It focuses on a virtual learning system (VLS). The theoretical background is mostly based on the Technology Acceptance Model (TAM), but extends the basic TAM with additional concepts from a different theory, in this case task-technology fit. It describes an empirical study in which 165 participants answered a questionnaire about a virtual learning system they have experience with. Several demographic concepts are elicited, but only used in a descriptive capacity to characterize the population. The variables from the data model – three TAM variables and one representing task-technology fit – are used for structural equation modelling. In the conclusion, the authors highlight the influence task-technology fit has on the standard TAM concepts. Similar articles, employing the same format but investigating other concepts (frequently in conjunction with TAM) and using other types of software system, were amongst the most frequent type included in the review.

4.2.5.1 Study level data

The general study data was intended for descriptive purposes and not used for answering the research questions in this chapter. It consisted of one record per article. Beyond bibliographic data, we recorded the type of software system on which the study was done and whether the study measured satisfaction for distinct features of the software features separately or only for the software system as a whole. We also had a note field for remarking what's especially interesting in the study, which was not included in the later evaluation but helped us orient ourselves in our dataset. Table 4.3 represents a sample record. In our dataset, we used a standard table format (each record in a row). The table shown here is transposed for better readability.

4.2.5.2 Variable level data

Each article which fulfilled our inclusion criteria describes an empirical study in which a data model was calculated. The data model consists of $N \geq 2$ variables, and relationships between them. At least one of these variables is *satisfaction*. The aim of this data extraction step was to create a list of all variables present in our primary sources, then map it to the concepts measured by the variables. The consolidated list of concepts provided an answer to RQ 6.

For each article, we extracted all variables of the data model described in it. Concepts which were mentioned in the theoretical section of an article but not measured with variables were not extracted. Similarly, variables which were measured but not included in the data model were not extracted. These were typically demographic variables, which were reported for descriptive purposes, but not evaluated in relation to satisfaction. In two cases [141, 116], demographic variables such as *age* were treated by the authors as part of the data model, and correlations with satisfaction reported. The demographic variables from these two studies were included in the dataset.

We extracted four data fields for each variable. The first field is the name for the variable given in the source article. The second field is a short reference string which serves as a foreign key to the study level data. It was followed by a field listing the definition given in the article for the concept behind the variable. This was done not only to gain a better description of the HCI concepts used in satisfaction measurement, but also to be used in the aggregation step reported in the next section. As discussed in the evidence quality section, less than half of all variables were defined, so we had to leave the field empty when a definition was missing.

The last field is a Boolean variable denoting whether the study reports a connection of the variable to satisfaction. It was set to *TRUE* when the article reported a measurement of the strength of the connection, regardless of whether it was included in an explicit representation of an idealized model or not. Thus, in articles listing a full correlation matrix, all variables were seen as connected to satisfaction, because we had data to evaluate the strength of these connections. To simplify the evaluation algorithm used in later stages, satisfaction was always assumed to be reflectively connected to itself.

To illustrate our approach, we continue the example of Lin's study on web learning performance introduced in the previous section [138]. The variable level data from this model was extracted in four records, as shown in table 4.4. The variable names stem from the data model representation, while the definition was copied from the background section of the article. The variables *perceived fit* and *VLS continuance intention* are connected to satisfaction, as evidenced in the graphical representation of the model. Satisfaction is by convention connected to itself. *Positive impacts on learning* has no connection to satisfaction shown in the model, but elsewhere in the article there is a full correlation matrix providing a measurement of that connection. A short reference column was included in the dataset but not shown here, as it only functions as a foreign key to the study level data.

Name	Author definition	Connected to satisfaction
perceived fit	–	TRUE
satisfaction	–	TRUE
VLS continuance intention	IS continuance behavior is defined as the continued sage of IS by adopters, where a continuance decision follows an initial acceptance decision	TRUE
positive impacts on learning	–	TRUE

Table 4.4: Variable level data example for a sample article [138].

4.2.5.3 Relationship level data

A data model operationalizes concepts and their relationships. In our review, we define a relationship as one which connects exactly two variables and reports a numeric measurement of relationship strength. While it is statistically possible to create data models which incorporate multivariable interactions, it is complex and only a very small number of our primary sources employed such techniques. To simplify our analysis, we ignored data on variable interactions and only extracted relationships between two variables.

The studies in our dataset use three different measurements of relationship strength, as determined by the inclusion criteria: *correlation coefficients*, *regression coefficients* and *path coefficients*. However, neither regression coefficients nor the related path coefficients are suitable for comparison across studies, as their value is always dependent on the choice of other variables used in the data model [45]. Thus, we could extract variable data from variables linked by any of the three strength measurements, but for the relationship data, we restricted the extraction to relationships whose strength was measured with a correlation coefficient.

The relationship definition we use does not include a relationship direction. This direction is frequently present in the primary sources, because authors assume causality in their theories. As these causal directions are neither relevant to our research questions, nor empirically proven, we did not extract direction data from our sources.

A quantitative metaanalysis involves calculating the variance of each measurement. For this, we included the sample size in our data extraction. In some cases, a publication reported multiple measurements of the same relationship in different samples (e.g. it reported multiple studies, or it reported a study with subgroup design). In these cases, we also extracted each relationship measurement as a distinct record.

The recommended procedure for data extraction is to have at least two coders and have each study examined at least twice. As no two coders were available, one researcher extracted all relationships to satisfaction a second time with a delay of several months. This enabled us to

correct errors in both the variables and the relationships extracted from each study. As relationships between two non-satisfaction concepts were only used for benchmark purposes, the effects of a small number of extraction errors among them are negligible. Therefore, we did not extract those a second time.

The record schema for our relationship level data included the name of the two variables as used in the primary study, a number for the sample (only entered if there was more than one sample in the publication), the sample size, a numerical value of the correlation coefficient, and a name of the satisfaction variable (only entered if there was more than one per sample).

A fully normalized data schema would have required separate extraction of sample level data, decreasing our manual data entry speed and necessitating more complex analysis code. We opted for a partly denormalized schema instead, necessitating the redundant entry of the sample size on the relationship level. Potential errors due to inconsistent entry are prevented by a check in the analysis code, which reports an error if different sizes have been entered for the same sample.

First variable	Second variable	Sample	Sample size	Correlation coefficient
satisfaction	VLS continuance intention	1	165	0.597
satisfaction	perceived fit	1	165	0.624
satisfaction	positive impacts on learning	1	165	0.561
VLS continuance intention	positive impacts on learning	1	165	0.654
perceived fit	positive impacts on learning	1	165	0.720
VLS continuance intention	perceived fit	1	165	0.741

Table 4.5: Relationship level data example for a sample article [138].

The relationship data extracted from the study in the running example is presented in table 4.5. Analogously to the variable level data, a column with a short string referencing the study, together with a column with the sample number, is not shown. These columns are only necessary for data integrity and not used in the analysis.

4.3 List of concepts connected to satisfaction

This section describes our work on research question 6. We started with the variable level data from our dataset and extracted the concepts described by the variables. We then filtered the concepts and categorized them using the ASMA schema. The resulting list represents the answer to the research question, an enumeration of concepts which have been investigated in relation to satisfaction.

4.3.1 Preparing the variable level data

Our variable level data consisted of 1102 records. While we extracted those, we noted that authors would frequently use non-standard names for variables. This means that many of the variables in our records referred to the same concept, but a naive comparison by name would have resulted as treating them like two separate concepts. In a few cases, we also noted a homonym problem, where two authors used the same name for two different concepts.

To alleviate this problem, we added a new field to our dataset, *standardized name*. It contained a new, more comparable name we chose for the concept behind the variable.

The purpose of renaming was to ensure comparability between variables and prepare the dataset for creating a concept-level list. In most cases, this meant choosing a new name such that two variables which describe the same concept would be mapped to the same standardized name. Sometimes, we also generalized variables, seeing them as a very concrete measure of a more general concept. We only mapped variables to a new name when we considered it necessary for our evaluation. For 364 variables, the original and standardized name were identical.

We used following rules for choosing a standardized name:

1. When the author used a long name, but a shorter name was sufficient for understandability. For example, the variable *end user computer satisfaction* [56] was mapped to *satisfaction*. We kept longer names if they were needed to distinguish from a similar, but not identical variable present in the dataset.

2. When the variable name was a linguistic variation of a concept name already present in the dataset. For example, in [21], we changed *self-efficacy* to *self efficacy*.

3. Some studies calculated relationships between numbered questionnaire items lacking a name. We mapped them to a concept based on the question text. For example, item 9 from [12] was measured on a Likert scale labelled "The Intranet is easy to use (e.g. personalization, handling the employee-directory)". We mapped it to *ease of use*.

4. When variables had names specific to the system type used in the study, we mapped them to a non system specific concept. Thus *learner's characteristics* from a study of an electronic learning system [4] was mapped to *user characteristics*.

5. In some cases, we generalized the variable used by the author. For example, *switching costs* from [28] was mapped to *cost*, even though from an economics point of view, switching costs are only one of many possible types of costs.

6. When a variable's name was synonymous with another name already present in the dataset. As an example, multiple authors used *emotion* as a variable, but Dzikovska et al. [57] used *affect*. We chose the name *emotion* for the concept behind both emotion and affect.

7. When two variables described the same concept even though their names were not synonymous. These decisions were made based on definitions provided by the authors. For example, Ogara et al. [151] use a variable called *user experience*, and state "We define user experience as the extent to which a user gains familiarity with mIM, which allows them to connect with communication partners". We mapped this variable to *familiarity*, as this would correspond to the familiarity variable used by other authors.

8. When a variable was the opposite of a variable already present in the dataset. A common example was *confirmation*, which was a term used by multiple authors. There is an influential satisfaction theory by Oliver [153], the expectation-disconfirmation theory, and many authors use the *disconfirmation* variable, while others use *confirmation* as the opposite. We are aware that in psychometric measurements, the negative and positive ends of the same scale will not always exhibit a symmetric effect on other variables. Still, as we needed a greater comparability in this review, we felt that the added imprecision is acceptable in this case.

9. If none of the above applied, the variable was mapped to a concept of the same name.

For comparison, we list the variables found in our example study and their mapping to concepts in table 4.6.

Variable	Concept	Rule	Rationale for mapping
Perceived fit	task technology fit	7	"Task technology fit" is the common name, as discussed in the text of the original article.
Satisfaction	satisfaction	9	Mapped to concept of same name
VLS Continuance intention	continuance intention	4	This is a frequently used variable, and in other studies, it does not refer to a virtual learning system.
Positive impacts on learning	task outcome	7	The variable name is too specific to this study and not comparable to others.

Table 4.6: Mapping of variables to concepts in the example study

Before the mapping step took place, our dataset consisted of 1102 variables, with 728 distinct names. After the mapping, the records contained 213 distinct standardized concept names.

4.3.2 Extracting and refining the concept level data

To answer our research questions, we needed to make statements on the concept level. After the variables were mapped to concepts, we compiled a concept list, filtered it, added definitions and assigned the concepts to ASMA categories.

The filtering was necessary to achieve the desired level of evidence. We considered a single study to be insufficient evidence, and required a second study to confirm the connection. Except for this condition, all found concepts were included in the further evaluation. Out of the 213 concepts, 80 met this condition. In the following, we are only working with the filtered concepts.

Table 4.7 contains the variables from the example study as they appear in the concept level dataset after preparation. It is transposed for readability. *Task technology fit* does not appear in the table, as it is not among the concepts included in the final selection.

4.3.3 Results

The data preparation step described in section 4.2.5 produced a list of 80 HCI concepts which have been investigated together with satisfaction, with added ASMA categories. We tabulated that list by category, as presented in table 4.8. This result answered our first research question of this chapter, RQ 6.

4.4 Strength of the relationship between satisfaction and HCI concepts

This analysis was needed to answer the second research question of the literature review. We prepared the available data and conducted a metaanalysis to derive the needed results.

Name	Satisfaction	Continuance intention	Task outcome
Occurrences	131	52	6
Defnition	The users' overall feelings based on their experience with the system	The user's decision to continue using the system over a long period of time	Denotes whether the user was able to complete the task successfully
Source	Lee at al. [127]	Bhattacherjee [20]	our definition
ASMA entity	user	user	context
ASMA dimension	mutable properties	activity	mutable properties

Table 4.7: Transformed data on the concept level. Concept shown from the example study by Lin [138].

	activity	stable properties	mutable properties	appraisal
user	– complaint – continuance intention – effort – habit – recommendation – use	– age – commitment – education level – intellect – loyalty – personal innovativeness – technology experience	– anxiety – attention – attitude – behavioral control – disconfirmation – emotion – enjoyment – expectation – flow – preparedness – satisfaction – self efficacy – trust	
system	– effectiveness – efficiency – error rate – speed	– accessibility – adaptability – assurance – comfort – complexity – credibility – empathy – familiarity – flexibility – interactivity – learnability – navigation – reliability – responsiveness – security – social presence		– aesthetics – design – ease of use – service quality – system quality – system rating – usability – usefulness
information		– accuracy – completeness – content – media richness – understandability	– currency – format – relevance – timeliness	– information quality
context		– alternatives – corporate image – cost – dynamic capability – management support – marketing – subjective norm – support – user involvement – user participation	– benefit – fairness – risk – social influence – task – task outcome – value	

Table 4.8: Concepts which have been linked to satisfaction

4.4.1 Preparation of the relationship level data and metaanalysis

In the data extraction step, we recorded the relationships between variables used in the primary studies and satisfction as described in section 4.2.5.3. As a first step towards answering our research question, we replaced the variables by the concepts behind them, as described in section 4.3.1. This resulted in concept-level relationships. As multiple variables can correspond to the same concept, and some authors also measured the relationship of the same concept to different forms of satisfaction, the resulting dataset could contain multiple measurements of the same concept-level relationship per sample. These measurements are not independent, so it is not permissible to use them as separate measurements in the metaanalysis [45]. Instead, we calculated an arithmetic mean of all repeated measurements per sample and used this as the sample measurement for that concept.

The primary studies rarely contained sufficient information to determine the direction of a measurement scale used for a variable. Thus, when we encountered a negative correlation coefficient between a given concept and satisfaction, we could not determine if it is due to an inverse relationship, or a scale defined in an unexpected direction. Some of the positive correlation coefficients are probably also calculated based on a scale going in a direction opposite of what the variable name would suggest. We therefore decided to not distinguish between positive and negative correlations and to only analyze the absolute values of the correlation coefficients. Our results therefore only reflect the strength, but not the direction of the relationship to satisfaction.

With the relationship data prepared, we proceeded to conduct a quantitative metaanalysis of relationship strength. The methods available for quantitative metaanalysis are imprecise when only a small number of samples is available. We therefore followed Higgins' recommendation [87] and only conducted the metaanalysis when at least 5 samples measured a relationship for the same concept to satisfaction, regardless of the number of variables measured per concept in a single study. This resulted in a list of 26 concepts we analysed in the next step.

For the metaanalysis, we chose to use a random effects model. Random effects models assume that the measurement variance between samples is influenced by differences in the context of measurement, while fixed effect models assume that the variance is only due to measurement error. As there is no standardized way of measuring satisfaction or its related concepts, we assume that conceptual differences between the primary studies are highly relevant, which makes a random effects model the better choice.

For the calculation itself, we used the R package *metafor* [186]. It calculates an estimate for the measured metric (in this case the correlation coefficient between satisfaction and a HCI concept), a confidence interval for the estimate, as well as measures of heterogeneity. While it provides a wide choice of estimators, we used the default setting of a restricted maximum likelihood estimator, which is approximately unbiased [186]. To avoid ceiling issues, we used a Z transformation [45] in the calculation of estimates, then converted the results back to correlation units for easier interpretation.

strength category	none	weak	medium	strong
center	0	0.2	0.37	0.53
determined as	lower endpoint	first quartile	median	third quartile

Table 4.9: A definition of categories of effect strength for HCI concepts. The first and last categories arise from the natural limits of a correlation coefficient's absolute value, while the middle three were derived from the sample distribution of non-satisfaction relationships in our dataset.

This resulted in a numerical estimate for the correlation coefficient. That estimate is of limited value for practice, as it is imprecise and needs to be interpreted. The recommended practice is to contrast estimated effect sizes to other known effect sizes from the same domain, and use this comparison to describe the effect's strength.

Such data was contained in our dataset in the form of relationships between two concepts other than satisfaction. We used this data as our benchmark. Specifically, we calculated the quartiles of the sample distribution of all relationships between two distinct non-satisfaction concepts. We then used a least squares method to categorize our results using the scale beginning (zero) and the quartiles as centroids. The upper scale endpoint was not used, as unlike raw correlations, the Z-transformed values do not have an upper bound. Table 4.9 shows the definition and numerical value of the strength category centers.

The simple assignment of estimated values to strength categories does not contain information on how certain it is that the real values behind them belong to the respective strength categories. To gain a better understanding for that, we conducted a one-sided hypothesis test. We tested H_1 *The correlation is larger than the center of the next-lowest category* against H_0, *The correlation is not different from the center of the next-lowest category*. The test was conducted on the Z transformed data, using the known approximations for sampling variance and sampling distribution for Z-transformed values. We conducted the test once for each concept and controlled the family wise error rate with Holm's method [90].

4.4.2 Results

From the 80 concepts we identified in the previous research question, 26 had at least 5 correlation measurements in our dataset. The strength of their relationships to satisfaction is summarized in table 4.10. A much more detailed description of the raw data and calculations for each concept is available in appendix A.

Strong relationship			
attitude	content*	continuance intention*	currency
disconfirmation*	ease of use*	efficiency	enjoyment*
format	information quality*	loyalty	reliability
service quality	support*	system quality*	task outcome
trust	usefulness*	value	
Medium relationship			
benefit	expectation	self efficacy	social influence*
Weak relationship			
technology experience*	use*	user participation	
Insufficient data for metaanalysis			
accessibility	accuracy	adaptability	aesthetics
age	alternatives	anxiety	assurance
attention	behavioral control	comfort	commitment
complaint	completeness	complexity	corporate image
cost	credibility	design	dynamic capability
education level	effort	emotion	empathy
error rate	fairness	familiarity	flexibility
flow	habit	intellect	interactivity
learnability	management support	marketing	media richness
navigation	personal innovativeness	preparedness	recommendation
relevance	responsiveness	risk	satisfaction
security	social presence	speed	subjective norm
system rating	task	timeliness	understandability
usability	user involvement		

Table 4.10: List of the HCI concepts linked to satisfaction, categorized by relationship strength. An asterisk denotes concepts whose strength is significantly different from the center of the next lower category ($\alpha = 0.05$).

4.5 Discussion

In this systematic literature review, we first created a list of all concepts connected to satisfaction in the literature we found. As a second step, we conducted a metaanalysis to determine the strength of these connections.

In this section, we give a brief commentary on these results. As we worked on multiple concepts at once, the results cannot be summarized into new insights. Rather, they represent a reference body which can be used by researchers or practitioners in their work in satisfaction measurement. In this thesis, the results are used in chapter 5, where we describe a method which employs the results of this chapter too create a questionnaire for anticipated satisfaction.

4.5.1 Concepts related to satisfaction

We found that there are several established models measuring user satisfaction, such as Delone and McLean's model for information system quality [51], the Doll and Torkzadeh instrument for end user computer satisfaction [56], and the technology acceptance model and its successor UTAUT [183], to name a few. Each of the established models uses only a few concepts. However, many authors use either a derivative version of these models, or a new model, which introduces a multitude of other concepts.

The number of distinct concepts we found is roughly 1.5 times the number of studies in our sample. Even after filtering out the concepts measured only a single time, there are 80 concepts which have been found to influence or be influenced by satisfaction, and they cover all entities present in the satisfaction measurement as evidenced by their ASMA categorization.

This finding shows that satisfaction cannot be explained with a simple idealized model. HCI concepts form a complex network and are highly interrelated. They are also very interdisciplinary. Some of the concepts we found are typical for software engineering, such as the *responsiveness* of a system, while others play a major role in other fields such as marketing, psychology, sociology and management science. This underscores the importance of satisfaction as a central concept with far reaching implications in HCI.

4.5.2 Strength of the relationship of the found concepts to satisfaction

For 26 concepts, we were able to do a metaanalysis, quantifying their relationship strength to satisfaction. It is interesting to see their distribution among the strength categories, and to pay special attention to outliers, commonly used concepts which turn out to have a weak relationship.

The metaanalysis results appear to be a convenient, firm number. In reality, they are a tentative estimate. Before they are used for comparing or choosing concepts, their uncertainty has to be taken into account. This uncertainty can be qualified by paying attention to problematic patterns in the raw data, such as heterogeneity measures and an estimate for missing data and its possible impact.

Strength categorization The HCI concepts in our dataset show a tendency for high correlations overall. The standard category centers used in absence of benchmarks are 0.1 for weak, 0.3 for medium and 0.5 for strong correlation, while our data yielded 0.2 for weak, 0.37 for medium and 0.53 for strong. This is probably due to the fact that many usability concepts are some form of evaluation (for example, *format* is not some quantitative measure of the information's format, but the participants' opinion of how good the format is) and thus subject to the users' tendency to use the positive half of a scale, and also that evaluations of usability concepts frequently overlap.

Even though the non-satisfaction correlations set a high benchmark, the satisfaction correlations are even higher. From the 26 analyzed concepts, none fall into the *no relationship* category. The lowest estimated correlation is 0.22, for the concept *user participation*. It is surprising to find that, despite that *user participation* has been found to contribute considerably to system quality [1], it does not correlate well with satisfaction. Additional research is needed to discover the causes for this result.

The other two concepts with weak correlation are *use* and *technology experience*. It is unsurprising that the users' experience with technology has little influence on their satisfaction with a specific system – both for simple and complex system, it is likely that the knowledge about the system itself is more important than knowledge about technology in general. In interpreting the low result for *use*, it should be noted that for many of the systems in the primary studies, the participants may have had no alternative to using the system (e.g. a business information system mandated by their employer, or an e-learning system used in a course they are taking) or have only had a choice between a few similar alternatives (e.g. mobile data networks in their country).

The medium category also holds very few concepts. While they are important for other HCI research, they are rarely seen as predictor of satisfaction. The direct relationship of expectation to satisfaction is especially interesting for the measurement of expected satisfaction, as it is the only strictly pre-exposure measurement on the list.

The majority of concepts (73%) fall into the *strong* category. It is notable that all concepts describing the entities *system* and *information* in the ASMA schema are in this category. It is probable that they have a more direct influence on satisfaction than other HCI concepts.

Heterogeneity measures The metaanalysis of all concepts exhibited high heterogeneity levels. We calculated the I^2 metric, which quantifies what percentage of the variance in the effect sizes is due to variance between the studies as opposed to sampling error [45]. With the exception of a single outlier, it ranged between 82% for *trust* and 99.18% for *format*. This is considerable variation which is not due to measuring errors, but to the differences in study design and the measurement context of the different studies.

From a theory-building point of view, the high heterogeneity is a sign that the measurements are sensitive to moderators. While the primary studies frequently use multivariate models to explain satisfaction, the inclusion of interaction effects in the models is very rare. Our results show that they are likely to be needed for the advancement of satisfaction research, as the direct relationships show a high variability.

For practitioners, the causes of heterogeneity and their exact effects are less important than heterogeneity's impact on measurement results. The main risk is that they may choose to use a concept which has a low effect on satisfaction in their measurement situation. While we do not have sufficient data to suggest instrument choice, our results are useful for determining which

concepts are likely to have a low worst-case performance, enabling a minmax-based decision process. The forest plots we provide in the appendix serve as a visual aid for that.

Missing data Systematically missing data due to publication bias creates the risk of calculating artificially high results in a metaanalysis [45]. A high risk factor is our data collection protocol, which only included peer reviewed publications. However, our primary studies do not focus on a single binary relationship, but use large data models with several variables and multiple relationships, frequently even publishing a full correlation matrix. Also, software engineering journals do not have firm requirements for statistical significance. Thus, even when some of the variables in a model have a correlation to satisfaction close to zero, the remaining results are sufficient for an article to be published. While we have very few raw correlation measurements close to zero, we think this is more due to true collinearity between HCI concepts and to a lesser degree due to a publication bias. The funnel plots of the data are difficult to interpret, due to the small number of datapoints and the high heterogeneity. Still, they do not exhibit the typical pattern seen in datasets censored by publication bias.

4.5.3 Strengths and weaknesses

For conducting the systematic literature review, we followed a textbook [45] to ensure high research quality. Still, our research is subject to the limitations inherent in empirical research.

Construct validity The central construct in our review is the HCI concept of *satisfaction*, and it is present in all primary studies we analyzed. As the goal of our study was to identify the different methods of measuring it, it was impossible to compare it to a selected "standard" method of measurement.

An important step in our method included mapping the variables found in the primary studies to HCI concepts. To ensure construct validity for these concepts, we tracked the definitions used in the primary studies and created a semantic mapping, bypassing the ambiguity inherent in variable names.

Internal validity As part of our study, we described 80 concepts explaining satisfaction. This creates a much larger model than usual, such that the risk of overlooking important explanatory variables is very low.

A potential problem for internal validity is that we were only able to consider direct relationships. Interactions and the influence of moderator variables were not part of our research.

External validity A study has external validity when its results are generalizable. A barrier to the generalizability of our results is the unfavorable ratio of variables to data points. This means that some of the patterns we are seeing are due to random noise rather than true effects, so that it cannot be concluded that they will hold for a larger population. On the positive side, the diversity of populations in the primary studies means that the results of our meta-analysis are applicable to broader populations than those of the primary studies alone.

Reliability Our structured method and the choice of suitable tools increase the reliability of our research. Nevertheless, we had to deviate from the guidelines by having a single researcher apply the inclusion and exclusion criteria and perform the data analysis.

An exhaustive search is difficult to achieve in such a large field, and having a single person perform it increases the chance of missing important studies. As a countermeasure, we used a precompiled list of important publications of the field of interest, and erred on the side of inclusiveness when selecting search results for further consideration.

To alleviate reliability problems, certain key steps were conducted twice by the author, with an interval of several months between the repeats. While this does not ensure a high interrater reliability, it prevents carelessness mistakes and emphasizes ambiguous decision, which can be reconsidered in more detail. As a second mitigation strategy, the doctoral supervisor was included in discussions of the study design, approved the methods used for the analysis and checked the results for plausibility.

We conducted our work to the standards of good scientific practice. Beside the reliability and validity factors discussed above, the work exhibits several strengths contributing to its quality.

Use of proven methodology Secondary research plays an important role in science, as it helps consolidate existing findings to arrive at more robust conclusions. However, creating a good methodology is not trivial, and the results are sensitive to methodological mistakes. We used state of the art methods as described in a textbook [45], supplemented with further literature on best practices in literature reviews and metaanalysis [112, 87]. This increases the validity of our results.

Uniqueness To our knowledge, this is the first study which aims to capture all concepts related to satisfaction. The closest similar study we are aware of is Mahmood's metaanalysis [143]. It only considers nine concepts, and it is based on studies published up to 1998, 16 years older than our newest data. Our study allows a comparison between the different concepts and shows which ones exhibit a consistently strong relationship to satisfaction over several varying measurement studies. This information is useful for both researchers and practitioners in the satisfaction measurement field, and it is not available from other sources.

Reproducible research We chose methods and tools specifically created to allow reproducible research. We documented our search and our data transformations and released our dataset. The statistical analysis was conducted using an R script, which we also released. Thus the analysis can be automatically repeated on this dataset or even on a new one. The textual reporting used knitr, a tool for automatic insertion of R expression in LaTeX code, which eliminates the possibility for errors to occur when transcribing the results to the thesis.

4.5.4 Conclusions

We conducted this review in the context of choosing concepts for satisfaction measurement. The results show that satisfaction measurement is very complex. It is impossible to suggest a small, well-defined model to measure in all cases. Instead, practitioners need to make a decision about the concepts to include in their measurement.

In many situations, using an existing instrument would be the right choice. These instruments are validated and widely used, allowing some comparability of results. However, there are situation where none of the existing instruments fits well enough. Then a pick-and-mix approach would be a better option. In that approach, the measurement is made with a newly designed questionnaire. The concepts in it are chosen according to the business needs and the measurement context in the project.

The measuring of anticipated satisfaction is a case where no firm instrument exists. The method we are developing for it employs a hybrid approach. In the next chapter we discover some concepts which are especially important for the measurement of anticipated satisfaction, and our method stipulates that these should be included in the questionnaire. They should be complemented by the use of concepts more generally connected to any type of satisfaction, and these have to be chosen from a suitable list, such as the one we developed in this chapter.

When choosing concepts for a questionnaire, it is useful to choose those which have a stronger relationship to satisfaction. For our list, we conducted a metaanalysis of correlations coefficients to provide the needed strength measure. However, we recommend to only use it as an additional criterion together with other considerations, not as the main basis for decisions. First, the dataset only contained sufficient information for calculating an estimate for 26 concepts, and the remaining ones cannot be ordered by this metric. Second, the estimates have a rather large uncertainty, with the 95% confidence interval of two concepts frequently overlapping, which means that the ordering based on the point estimates has a high likelihood of being wrong for a randomly chosen pair of concepts. Third, the very high heterogeneity measures suggest that in a given context measurement, the true value may be far off from the point estimate.

Despite these drawbacks, the estimate information is still useful for decision making. The more distance there is between the estimate for two concepts, the more likely that ordering them by estimate is correct – for example, a concept from the category *strong* is quite certain to perform

better than a concept from *medium*, while for two concepts from *medium*, it is not sure that the one with the higher point estimate will perform better in a measurement. This is why we suggest to not discard the information, but use it together with other considerations. A combination of research goals, high distinction between the selected concepts, and a strong relationship to satisfaction will yield a balanced and useful questionnaire.

5 Guidelines for measuring anticipated satisfaction

This chapter describes a method for measuring anticipated satisfaction, based on the findings in the previous chapters. It is written as a guideline for a person who wishes to utilize the method to measure anticipated satisfaction. We call the application of the method a *measurement project*, short for anticipated satisfaction measurement project. It is conducted in parallel with a *development project*, short for software development project.

The method we describe uses a survey-based approach. It is derived from the generic guidelines for surveys presented in Gray [73]. The three subsections of this chapter correspond to the three method stages – preparing, conducting the survey and evaluating the results.

This chapter only describes the activities specific to the measurement method in detail. Other activities which are common to all questionnaire-based methods, such as the distribution of questionnaires to participants, fall outside the scope of this dissertation. As requirements engineers are not usually educated in methods for conducting surveys, they may need to acquire the basics of that knowledge from other sources. Guidelines for planing and executing the basic stages of a survey can be found in the relevant literature, for example [164].

5.1 Stage 1: Survey design and preliminary planning

This is an extensive stage. We have adapted the activities suggested by Gray and show the process in a diagram in figure 5.1.

Research questions To ensure naming consistent with activity diagram naming conventions, we renamed this as *Formulate research questions*. While we intend our method to be primarily used outside of academia, surveys are still considered research when used in a commercial context, as seen in terms like “marketing research”. Thus we kept the term “research” as suggested by Gray.

Decide on information needed The “information needed” consists in our method of three types of measurements (discrepancy, satisfaction and demographic measurements) and three types of objects which are measured (individual features, data records and the whole

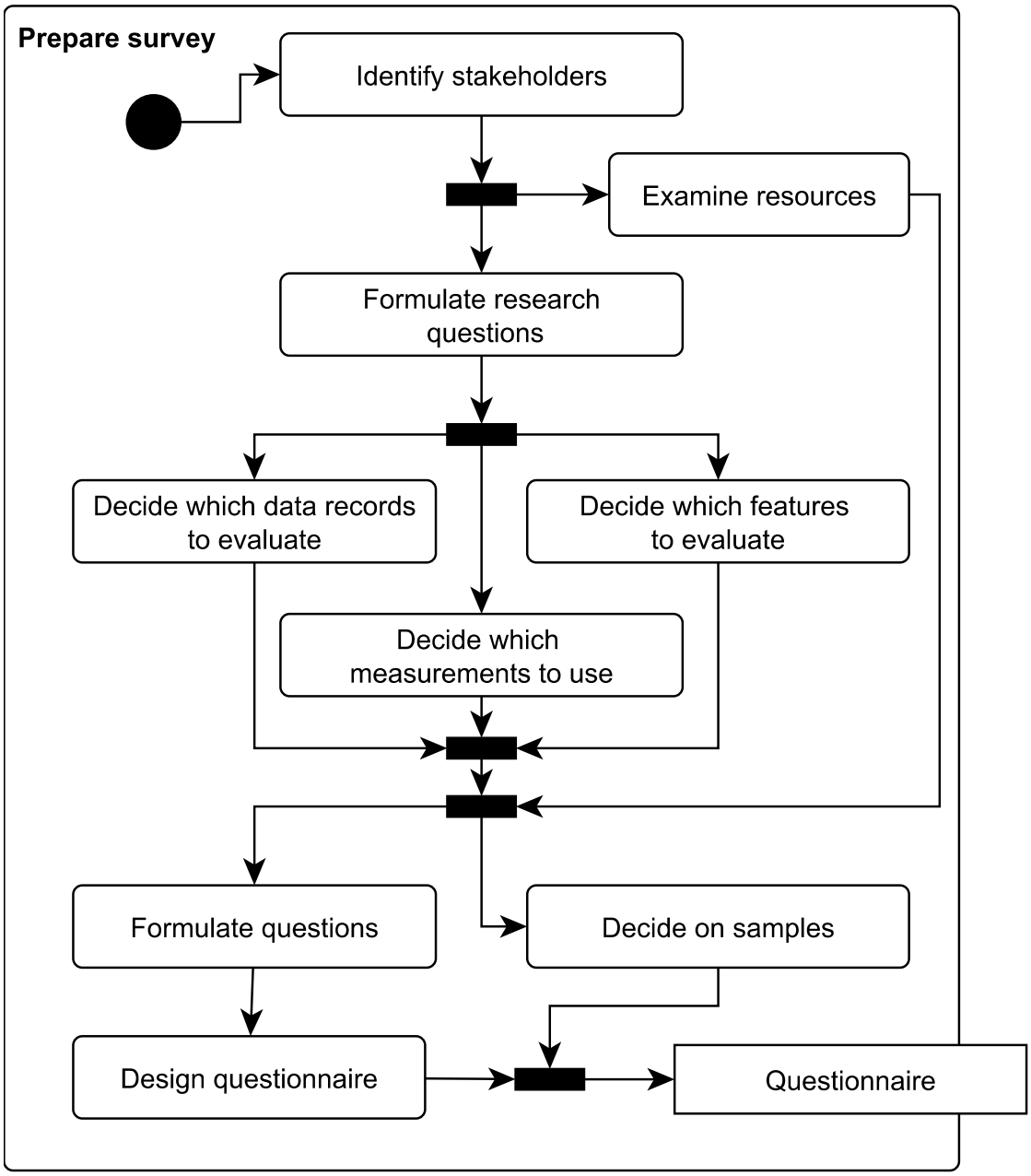


Figure 5.1: Activity diagram for *Prepare survey*. Adapted from Gray [73]

system). There is no need to make a decision about the discrepancy measures. This is because our model of anticipated satisfaction (developed in chapter 3) contains only three discrepancy measures, and all of them are used in the questionnaire. The system is seen as given in our measurement context, so it does not require any decision either. Thus, we split this into the three subactivities *Decide which data records to evaluate*, *Decide which features to evaluate* and *Decide which measurements to use*, containing both satisfaction and demographic measurements.

Review existing information on topic This activity is only necessary when designing a method from scratch. It is not applicable to our method.

Examine resources Adopted as-is

Decide on preliminary analysis approach and sample Our method employs an interview for the preliminary analysis. The activity is combined wiht the next one into *Decide on pilot sample*.

Decide sample See above

Choose survey method The word *method* is used on a different granularity level in Gray than in our research. It denotes the choice between interview, questionnaire, or other approaches for data collection. Our method for measuring anticipated satisfaction employs a questionnaire on this step, so the activity is omitted. The only decision is between using an online questionnaire, a paper questionnaire, or both, and this is usually predetermined by the target population's availability and preferences.

Structure and wording Changed to comply with activity diagram naming conventions. The new designation is *Formulate questionnaire content*.

Choose data processing method Again, the granularity of the term *method* is inconsistent with our usage. For us, it is an approach. Our satisfaction measurement method describes the data processing approach at length, so this step is not necessary.

Design questionnaire Adopted as-is.

In our experience, the process also requires an additional activity which has not been explicitly mentioned in Gray's list:

Identify stakeholders While the stakeholders of the development project are known, the stakeholders in the measurement project will differ somewhat, although there will be an overlap. We list the roles which are important to this project, and it is critical for the measurement project's success to identify the persons filling these roles.

Due to interdependencies of these activities, the list is roughly chronologically ordered. In practice, there will be some overlap such that activities are performed in parallel. There are also likely to be some iteration loops, with earlier activities being revisited and their results updated.

5.1.1 Identify stakeholders

This activity should be performed at the very beginning, as it can influence subsequent activity, including the activity of choosing a research question. It describes the roles of all participants in

the measurement project.

The description in this chapter assumes an arbitrary number of persons in each role. In reality, some roles will include multiple individuals (e.g. *users*), while others are likely to only have one person within the measurement project, for example a single *requirements engineer*. For grammatical simplicity, the terms for all roles will be used in plural only.

5.1.1.1 Requirements engineers

As described in the introduction chapter, we develop our method as a form of validation of software requirements. This falls under the responsibilities of a requirements engineer, so in the further text, we name the person who applies the method *requirements engineer*. This does not have to be congruent to the person's job description. It should be understood as a role within the measurement project.

The requirements engineers have the central role in the measurement project. They do the bulk of the work and are responsible for delivering the results.

5.1.1.2 Users

The research sample for the survey consists of prospective users of the measured software system. Therefore, the participants in the survey are refered to as *users*.

The choice of participants is discussed in more detail in section 5.1.6.2.

5.1.1.3 Test users

Test users are the ones who participate in the pilot study. Ideally, they are a subset of the users. If low user availability makes this not feasible, other participants can be recruited. Details on their choice is described in section 5.1.6.

5.1.1.4 Recipients

This is an umbrella term for the intended recipients of the survey results. In a typical development project, this will include the requirements engineers themselves, who may change the requirements specification based on the survey results, the project managers, who need metrics to judge the success of the development project, as well as managers from the customers' organisation, who wish a confirmation that the software system will fit their needs.

The customers' role in the software development project influences the choice of research questions and also of results representation. It is possible that the requirements engineers have to prepare different result documents for different customer groups, e.g. a detailed report for a product owner and a summary for a presentation in front of senior management.

5.1.1.5 Champions

The requirements engineers and the customers rarely have the authority over all resources needed for a measurement project. The project can need the sponsorship of influential individuals who have the authority to assign resources to the measurement project and have the experience and influence needed to champion the project in front of any relevant decision makers, e.g. the board of directors of the organization at which the users are employed.

Depending on the availability of resources, a champion may be or not be needed in a measurement project.

5.1.1.6 Revisors

Research, especially research with human subjects, is frequently subject to a host of legal regulations. There are bodies (e.g. committees) and officials who are responsible for verifying that all projects adhere to these regulations. In this description, *revisors* is used as an umbrella term for them. They can include for example the privacy officer and the human resources department of an organization at which the users are employed, or an ethic commission board.

5.1.2 Formulate research questions

The method measures different variables connected to user satisfaction for different features of the system. It is suitable for answering several different research questions.

The simplest case would be a summative usability study, which simply reports the level of satisfaction measured for each feature. However, this is not very informative, since predicted satisfaction is not an accurate prognosis of actual satisfaction. Also, we see no reason for a summative usability study before the system has been implemented. Instead, we assume that our method will be used for a formative usability study, and its results will inform decisions during the development process.

Below is a list of topics which can be investigated with the method. The list is not exhaustive, it just covers several broad issues which are worth investigating in a typical software development project.

Evaluation of individual features Does a feature seem problematic to the users, or will they be satisfied with it?

A/B test of features Because the method scales to a large amount of users, it's possible to make variations describing different ways of implementing a feature and compare the results for them

Problem magnitude Sometimes, due to a tradeoff between different stakeholders' constraints, a requirements specification will contain features which are not user-friendly, or which serve one group of users but impede another. The information gathered in the survey can serve as a prediction of the magnitude of the negative effect of a planned feature known to be unpopular.

The choice of a research question will influence the further preparations, especially the choice of features to be covered in the questionnaire. It is possible to answer multiple research questions with a single survey. This will depend on the nature of the questions, and the resource constraints, especially the number of users available and the number of survey questions which can be included in a questionnaire, as discussed in section 5.1.4. More details on the choice of features to evaluate are given in the next section.

The recipients should be involved in choosing the research questions. As a minimum, they should approve the questions created by the usability researchers.

5.1.3 Decide which features to evaluate

The length of the questionnaire restricts not only the number of concepts which can be measured, but also the number of features which can be evaluated with these measurements. The requirements engineers and the recipients have to create a list of features to be included in the survey.

As with measurement concepts, the choice of features will be determined by the largest possible contribution for the survey's goals. In our own applications of the method, we found following types of features to produce useful results:

Uncommon features These are innovative or highly specialized features which are rarely seen in other software systems. As a feature shared with other products cannot represent a competitive advantage, the unique features of the system play an important role in product success. They are also difficult to design well, as there are not many known examples of good implementations. It is less certain how users will react to them. Getting user feedback on this type of feature is more informative than getting user feedback on a very common feature such as "print a document". There is also a restriction here: such features are harder for users to imagine and can deliver less accurate results. If the feature's complexity

does not permit to describe it well in the limited space available in the questionnaire, it is not a good candidate even if it is considered important for product success.

Focal points of conflicting viewpoints The constraints and stakeholder goals in a development project can be in conflict with each other. A classic example is the conflict between security and usability goals, where users desire simple interfaces but are frequently required to comply with cumbersome security measures. Measuring the satisfaction with such a feature will indicate whether users are willing to accept the chosen tradeoff.

Good enough features Sometimes, the optimal way to design a feature is known but expensive to implement, so the system designers decide to specify a suboptimal version. A measurement of this feature can deliver evidence that the feature is or is not good enough as planned.

User-oriented features not proposed by users Requirements are elicited from different sources. It can happen that a source other than the users themselves proposes the inclusion of a feature whose sole purpose is to benefit the users. The measurement of anticipated satisfaction can be used to verify that such a feature is indeed desired by the users.

5.1.3.1 Decide which data records to evaluate

Beside questions on system functionality (represented as features), the questionnaire will contain questions on information (represented as data records). Anticipated satisfaction measurements pertaining to the *information* entity in the ASMA model are based on descriptions of data records provided in the questionnaire.

The types of data records to include in the questionnaire depends on the focus of the survey and the system under evaluation. If the recipients are certain that the data structure chosen is ideal for the users' needs (for example because the process in which this data is used is well established and such records have been in use for a long time), this part can be skipped altogether. If data records are included, there are likely to be fewer data records than features. For example, in our RefsQ study described in chapter 3.1 we evaluated a system for managing expenses. We had a single data record - the receipt received when incurring some expense - and 16 features.

5.1.4 Decide which measurements to use

5.1.4.1 Discrepancy measurements

The use of the method described here alreay predetermines the information which will be gathered: it will be information on anticipated satisfaction. Chapter 3 lists the concepts we found to be relevant for measurements of anticipated satisfaction in particular. The ones which can

be measured pre-exposure are *perceived understanding*, *familiarity* and *emotional attachment*. These concepts, which have consequences on the prediction's reliability, should be present in the questionnaire.

5.1.4.2 Satisfaction measurements

In our early studies described in chapter 3, we used a single concept to represent the set of *satisfaction-related concepts*. This choice was made to allow for more efficient research. For the actual method application, we recommend measuring multiple satisfaction-related concepts, in order to gain better understanding of the potential sources of (dis-)satisfaction.

We created a list of such 80 concepts from a systematic literature review, as described in chapter 4. Because the length of the questionnaire should be limited as to not overwhelm the participants, it is not feasible to use the full list. The requirements engineers have to select a subset of these concepts best suited to addressing most important issues in their development project.

The first consideration is to achieve a broad coverage of all satisfaction related concepts. Our recommendation is to choose a list of concepts such that all categories of the ASMA model are represented, as a good compromise between coverage and length. This is not a hard requirement. The list can be made shorter if the recipients find an incomplete coverage acceptable, or if they are interested in exploring a certain category in depth.

The second consideration is to obtain the most interesting information with a questionnaire of limited length. "Most interesting" will differ between development projects and will be determined by the research questions pursued with the measurement project as well as additional information available in the project. For example, if it is known that at least some potential users view a specific feature as "useless" (this information could have been noted during interviews with key users conducted for the purpose of requirements elicitation), the questionnaire could measure the concept of *usefulness* for this feature, in order to discover how prevalent the attitude is among a broader sample of users.

Everything else being equal, we recommend choosing concepts which exhibit a stronger relationship to satisfaction. Our metaanalysis reported in chapter 4 provides an estimate for relationship strength for the most commonly used concepts we found. Appendix A provides these estimates, as well as forest plots of the raw findings for each of the remaining concepts. We suggest that the requirements engineers consult this data when making a choice of concepts for the questionnaire. However, this estimate is not suited as the only criterion to choose concepts, as discussed in chapter 4.5.4.

As the questionnaire gathers information both on the level of the whole system and on the level of individual features, it is possible to tailor the selected concepts such that some of them are only applied to some of the features, while others can be applied to the whole system only.

5.1.4.3 Demographic measurements

Aside from the questions intended for measuring anticipated satisfaction, the requirements engineers can add demographic questions. These questions have two uses. First, the requirements engineers might wish to verify that the users who participate in the study are indeed part of the target population. Second, depending on the exact research question, it may be necessary to establish anticipated satisfaction within different segments of the target population. In this case, demographic questions are used to determine which segment the user falls in. Example for concepts measured with demographic questions would be the *age* or *job title*.

5.1.5 Examine resources

The reseources needed for the measurement project are similar to the resources needed in any survey method. For example, the requirements engineers need tools for creating the questionnaire, for storing the data, for statistical evaluation, for creating and maintaing the project documentation, etc. These are not described in detail here.

In our studies, the most frequent bottlenecks were access to users and the availability of feature descriptions. As the measurement project's target population is defined as the people who will use the system under evaluation, it is rarely possible to draw a sample from the general public. If the system is intended to be sold off the shelf, the potential users have to be identified. They do not have any comittment to the project and will rarely respond to the survey without good incentives. For bespoke systems, the future users are the employees of the organisation which orders the system. In this case, there are likely to be bureaucratic hurdles to conducting a survey among all employees.

The availability of feature descriptions was our second most problematic resource. Development projects employ a wide variety of approaches for requirements documentation. Sometimes, there are no written requirements available, or they do not contain a list of features which can be used in a questionnaire. In this case, a list has to be created specifically for the measurement project, as described in section 5.1.7.6.

5.1.6 Decide on samples

As the requirements engineers are not designing their method from scratch, the preliminary analysis does not need to be as extensive as in other survey projects. However, as there is no standardized questionnaire to use, we still recommend that they run a small pilot study to verify that the questionnaire can be answered as intended and that it isn't confusing users.

5.1.6.1 Pilot sample

In our own studies, we typically use 4-5 people to answer the questionnaire in the pilot stage. As the resulting questionnaires proved to be adequate for their purpose, we regard this number as sufficient. Ideally, they would be drawn from the target population. If this is not possible, they should share as much of the domain experience of the target population as possible. Also, it is important that they are not members of the development project. The reason is that a shared vocabulary usually evolves within a development project, which will be understandable to its members but cryptic to outsiders. While it is valuable to gather feedback from team members too, it is necessary to test the questionnaire on persons not accustomed to the internal vocabulary.

We instruct the participants of the pilot study to not only answer the questionnaire, but also to record any mistakes they found and to point out any formulations they found confusing. They also document the total time they needed, and give any other feedback they feel necessary. A short interview proved the easiest method for this. With a low number of pilot participants, it is feasible to interview all of them in person.

The pilot study usually results in improvements to the questionnaire. The data points are not sufficient for a preliminary evaluation. Such an evaluation can be done in order to test the tools prepared for it and the process involved, but the final numbers cannot be used to draw conclusions for the result of the larger study.

5.1.6.2 Decide sample

The decision on a sample requires three elements: A well-defined target population, a sample size and a sampling frame.

The target population in the method for measuring anticipated satisfaction consists of the users of the system under evaluation. Whether all users are equally relevant or only a segment of them depends on the exact research question.

The choice of sample size and sampling frame are not specific to our method and can be handled according to generic questionnaire construction guidelines.

5.1.7 Formulate questions

After the usability researchers have decided on the concepts to be measured in the survey, they have to operationalize them by creating a questionnaire with concrete questions. The questionnaire used for our method consists of following parts:

1. Introduction

2. Questions about user stable properties and demographic questions
3. Questions about data records, including data record representations
4. Questions about individual features, including feature descriptions
5. Questions about the system as a whole
6. Invitation for feedback and concluding words

This section describes how to construct each part of the questionnaire. For the accuracy-related concepts we developed in chapter 3, we suggest a standard wording we used throughout our studies. The satisfaction measuring concepts from the literature review do not have a standard operationalization, and we have not tested all of them. We suggest some rules for constructing questions for them below. We also provide questionnaires we used in our own studies in Appendix A, which can be used as an example for question wording.

Table 5.1 gives a quick overview over the content of each part of the questionnaire.

Part number	**Evaluated object**	**Measurements to use**
1	–	No measurements, just information to introduce the users to the study
2	User	Demographics, ASMA user stable properties
3	Data records	ASMA information stable properties, ASMA information mutable properties, ASMA information appraisal
4	Individual features	ASMA system stable properties, ASMA system appraisal, ASMA user mutable properties
5	Whole system	ASMA system stable properties, ASMA system appraisal, ASMA user mutable properties, ASMA context stable properties, ASMA context mutable properties
6	Anything related to the study	Open question only

Table 5.1: Questionnaire structure

5.1.7.1 Question wording

We have not developed a standardized questionnaire for measuring anticipated satisfaction. Instead, we suggest a flexible approach, in which the requirements engineers measure the concepts

Concept	Question about	Wording	Scale anchors		
Perceived understandability	Features	My conception of the way this feature will be implemented is	Clear	Vague	Non-existent
Familiarity	System	Have you worked with other <type of system> before? If yes, please enter the name in the comment field.	Offers a choice between "Yes" and "No" and has a small free text field labelled "comment"		
Emotion	System	How do you feel about the system?	Like	Be indifferent	Dislike

Table 5.2: Questions for the three discrepancy measurements. We use a 5 point scale, but only the two end points and the middle point are labelled. <name of data record> is a placeholder to be replaced with the name used in a description of the data record. <type of system> is a placeholder describing the evaluated system in a way users will understand, such as "e-mail client" or "expenses management system".

best suited to their goals. This causes the issue of choosing a formulation with which to measure the variable representing each concept [1].

We propose three approaches to arrive at such a formulation. In later subsections, we suggest approaches to choose for which part of the questionnaire.

Use our formulation We recommend this primarily for the three discrepancy measures, whose use we have validated in multiple studies. The requirements engineers can also look up the exact formulation we used for satisfaction measurements in our questionnaires and adopt it. As we used direct questions for this purpose, this will likely produce very similar results to asking a direct quesiton.

The questions we use for discrepancy measures are listed in table 5.2.

Ask a direct question This is our recommendation for the satisfaction measurement questions. When a variable is chosen, the requirements engineers can use its name to form a question. For example, for the variable of *precision* we used the formulation

> Is the animal line fact sheet at the beginning of the page a precise description of an animal line?

where an *animal line fact sheet* is a type of data record processed by the system and printed above the question, and an *animal line* is a domain specific term well known to the users.

[1] For an explanation of our usage of the terms *concept* and *variable*, see section 2.3.

Use formulation from literature The variables we suggest for satisfaction measurement come from existing literature. As one of our inclusion criteria for the literature review was that a study should contain an empirical validation for its idealized model, all studies in our review have measured the variables extracted from their models. In many of the articles we collected, the exact formulation used for each variable is given. If not available, this formulation should be obtainable e.g. by contacting the authors of the study.

Two issues have to be considered when choosing this approach. First, the operationalizations of the concepts into variables are not standardized. Different studies measure them using different question formulations. As a result, it is not really possible to choose one "correct" way to measure a variable from the literature. Second, a concept can frequently be measured using multiple questions. This is a sound approach, but it increases the questionnaire length. It also needs more extensive validation to prove that the questions are indeed measuring the underlying concept.

Appendix A gives a detailed list of all concepts we found in the literature study. Among other information, it also references all studies which measure a variable for a given concept. This can be used as a starting point to finding existing formulations.

5.1.7.2 Answer options wording

The scalability of our method is achieved through the use of closed questions. The usability researchers have to provide answer options along with the questions. We recommend using a single type of answer scale for all accuracy and satisfaction measurements to ensure consistency. Exceptions can be made where the question cannot be formulated in a way that the suggested scale type makes sense.

The scale we use in our own research is a five point symmetric scale. We anchor the two end points such that the rightmost one indicates that the variable is true for the object measured in the question (e.g. a feature) and the leftmost indicates that the opposite of the variable is true. For example, for the question on *usefulness*

> How useful will the system be for your work?

we used the rightmost point *Very useful* and the leftmost point *Not useful*.

We intentionally choose to use a scale with an odd number of items. We feel that a neutral measurement is valid for the type of variables our questionnaire measures. Forcing the users to always choose a positive or negative measurement on an even-numbered scale is easier to interpret, but hides the true values of the data when participants have no positive or negative preference. We also label the middle option to make it more clear how it relates to the endpoints of the scale. In the question on usefulness, we used the midpoint label *Somewhat useful*. When there is no obvious neutral word to label the middle choice, we frequently use the formulation

somewhat, followed by the positive end of the scale.

Typical scale lengths in usability questionnaires are 5 and 7 point scales. While some criticism exists on the shorter lengths [64], we chose to use 5 point scales in order to reduce the cognitive load per question, as the questionnaire as a whole is quite long. We did not observe interpolations in our paper based questionnaires.

For typographic and linguistic reasons, we do not label points between the middle and the two ends of the scale. We did not observe any problems resulting from this choice.

The requirements engineers can choose to use a different type of answer scales, such as even numbered scales or Likert scales. However, we have only validated the type of scale described here.

5.1.7.3 Content of the part *Introduction*

This is a short text at the beginning of the questionnaire. Its purpose is to inform the users of the purpose of the survey, and to give details pertinent to the survey organisation, e.g. how to contact the requirements engineers if they have questions. The content is not specific to the method, and so it can be written according to standard guidelines [73], [164], [177].

5.1.7.4 Content of the part *Questions about user stable properties and demographic questions*

As explained in section 5.1.4, the questionnaire can include demographic questions. This is a type of question used in many surveys, and somewhat expected for the users. In interviews following pilot studies, our test users commented that putting these questions first is preferable to them, as the familiarity of the questions helps them ease into the tasks, and also they like to get over what they see as a "preliminary" before they come to the questions specific to the study.

The answer options for the demographic questions will be predetermined by their purpose. For example, if the goal of the question is segmenting the target population by age, there answers have to describe age ranges. A job title question might need a free-text answering field, or list the job titles which constitute the target population. Therefore, the answer structure cannot be made consistent with the answering schema of the satisfaction measuring questions, which is described in the next subsection. Our test users did not report this inconsistency as an obstacle.

We also recommend adding any measurements of the category *user stable properties* in this part of the questionnaire. The reason is that demographic variables are also user stable properties, so they naturally fit together, creating a cohesive questionnaire part which can be answered at once, before the user moves on to a description of the system in the second part.

5.1.7.5 Content of the part *Questions about data records*

The functionality of the system consists of operations done on the data. As a consequence, the feature descriptions frequently contain references to data records. The relevant data records can be represented in a questionnaire in a much more concrete way than features, so placing the part on data records first allows the user to build a better understanding. For that purpose, the data questions make up the second part of the questionnaire, directly after the system-independent questions of user stable properties and demographics.

Before the description of data record begins, the users have to be given some information on the system as a whole. Else they are lacking context and their understanding is diminished. The part should begin with a short introduction of the system. It should explain which of the users' tasks will be supported by the system, and how the system will be integrated in the users' work process. This introduction should be targeted at the sample population's domain knowledge, and kept as short as possible. We also found that test users feel reassured when they are instructed to answer the following questions to the best of their understanding, even though they have not seen the system, and so we recommend including such a notice below the system introduction.

The remainder of this part is a list of data records with questions about each data record. This requires the usability researcher to formulate three types of artefact: a data record representation, a question (one for each variable being measured), and an answer scale.

Data record representation The purpose of the questionnaire is to evaluate the content of the information independently of the final layout and graphic design which will be used in the system. To achieve this, the questionnaire uses either list of data fields and example values presented in tabular form, or a low-fidelity mockup reduced to only data fields, without buttons or other interactive elements beside fields for data entry.

We found that providing example data enhances the users' understanding. The test users reacted best to example data describing a complete real-life entity, as opposed to a record filled with artificially devised test data.

If the system processes user-generated information, the data record representation should indicate which fields are to be filled by the user and which are already available or calculated by the system.

Questions of the *information* category can relate to the record as a whole, or require an answer for each of the fields. This depends on the knowledge desired by the customers. If an answer is required for each field, it is recommended to add a clear numbering to each field, which can be referenced in a question's answer matrix.

Questions We do not use the three standard discrepancy questions for data records, for following reasons:

Perceived understandability This is necessary in features, where we want to know if the users have understood the description sufficiently to imagine the features. Examples of the data records exist before the system has been developed and can be reproduced in the questionnaire, thus no imagination is required. Instead, we measure the *actual understandability* of data. We use a direct question for that.

Familiarity In the studies we conducted, we already knew how familiar the users are with the data records, and measuring it would not have brought any new information. For example, in the measurement of the expense application, we assumed that everyone is familiar with a receipt. In the measurement of the application for tumor models, we knew that we are using a novel format, and that no user will ever have encountered it.

If the requirements engineers are not aware of how familiar the users are with the data records used, and especially if they suspect that there is large variation between users, they can include a familiarity question analogous to the familiarity question for features.

Emotion In our experience, people can have strong emotions about whole systems, as evidenced by e.g. flame wars on Internet discussion forums. We have no evidence that people experience such strong emotions about the structure of a data record. Thus we feel that such a question is unnecessary in this part of the questionnaire.

From the satisfaction measurement questions, we use questions describing the *information* entity. In our own studies, we used a single direct question per variable. However, the variables from the information-related categories are known to have validated operationalizations, for example instruments based on the popular Doll-Torkzadeh model [55]. If a variable is especially important to the customers, it can be measured with one of these operationalizations instead.

5.1.7.6 Content of the part *Questions about individual features*

Feature description As reported in chapter 3, the exact format of a feature description does not influence the understandability of the questionnaire or the result of the satisfaction measurement, at least for the three formats we tested (user tasks, user stories and sentence templates). We assume that in a development project, some form of requirements specification exists, and that it contains a list of features. Our recommendation is to use the feature description from this specification.

Alternatively, the requirements engineers can write new feature descriptions specifically for the questionnaire. This option consumes more resources and lengthens the preparation time, but it may result in descriptions better suited for the purposes of the questionnaire. We recommend it if the existing descriptions are found to have insufficient information, to not be understandable

when taken out of the context of the specification document, or to be too lengthy for inclusion in the questionnaire. It is also the only available approach if there are no descriptions available.

Feature descriptions can be written on different levels of granularity. In our studies, we started from a requirements specification in the form of user tasks as described by Lauesen [123]. We regarded each solution to a subtask as a feature. This led to features which the user perceives as distinct system functions, for example printing a report. The method has not been validated for other levels of granularity.

Questions For each feature, the questionnaire should contain a question on *perceived understanding*. We did not elicit data on *emotion* and *familiarity* in our studies, but from the free text feedback we concluded that users can certainly show emotions about a single feature, not just about the whole system. It is also possible that they are not very familiar with a system as a whole, but are familiar with individual features which they have used in other system types. So, if the questionnaire length permits it, taking these measurements could improve the data quality. These three discrepancy measures can be measured with the questions we formulated for them, as described in section 5.1.7.1.

The satisfaction measurements for the features are chosen from the concepts describing the categories *system stable properties*, *system appraisal* and *user mutable properties* in ASMA, where the two *system* categories contain concepts referring directly to the system, such as *complexity*, and the *user* related concepts usually describe some part of the user's cognitive or affective state which will change in response to the interaction with the system, such as *enjoyment*. They can be best measured with direct questions, but choosing questions from the literature is also an option.

5.1.7.7 Content of the part *Questions about the system as a whole*

We elicit all three discrepancy measures for the whole system, since each of them has an effect on the outcome. Examples can be found in the questionnaire texts in Appendix B.

The satisfaction measurement variables for the system cover the same categories as the variables for the individual features. Additionally, questions about the *context* entity can be included here. We do not envision many research questions which justify the elicitation of context measurements, so we have not used them in our own questionnaires.

5.1.7.8 Content of the part *Invitation for feedback*

The questionnaire concludes with an invitation for feedback, pointing out that we welcome any kind of comments on the system, the study itself, and others. Many users choose to not fill anything into that field, but the few comments which come back frequently contain valuable

information which can be used for improvements in the system or in the organization of future surveys.

Our test users had a positive reaction to this field, remarking that it promotes a positive attitude towards the measurement project and by extension towards the development project.

5.1.7.9 Questionnaire length

The questionnaire length is restricted by the time users can invest in answering it.

In the simplest case, the users are willing to fill out a questionnaire of any length. Then the limit is caused by their ability of sustaining focused attention on the task without reduction in performance. From our experience, we see 90 minutes as the absolute upper limit of questionnaire fill out duration, but recommend much shorter times if possible. The questionnaire design literature suggests that significant fatigue sets in after 30 minutes, but that drop our rate is also dependent on questionnaire quality besides length [23].

In reality, the availability of users frequently depends on the questionnaire length. If the users are professionals compensated for their time at market rates, long questionnaires make the measurement project prohibitively expensive. If the users volunteer their time, the response rate is inversely correlated with questionnaire length. The requirements engineers have to take into account the expected response rate, the size of the participant pool from which they can draw users, as well as other constraints of organisation in order to decide on an acceptable length of the questionnaire, represented as the average time users need to fill it out.

Once the time frame for filling out has been chosen, the requirements engineers can use it to approximate the number of questions they can include. In our studies, respondents needed between 20 and 60 seconds per question. Our questionnaire uses multiple similarly structured questions, often repeating the same question for different features or data records, thus making individual questions quick to process.

Estimating the exact length of a questionnaire is difficult. It is necessary to record the time the test users need in the pilot study, or at least to tell them the targeted duration interval and ask them to record if they fell within it. This information can be used to make a final questions selection for the actual study.

5.1.8 Design questionnaire

After the questionnaire content has been created as described in section 5.1.7, the design activity is a purely typesetting task. Its details fall outside of the scope of this thesis.

5.2 Stage 2: Conduct survey

Conducting the survey has three steps – piloting the questionnaire, amending it based on the insights from the pilot, then moving on to the actual survey. Figure 5.2 depicts the process as an activity diagram.

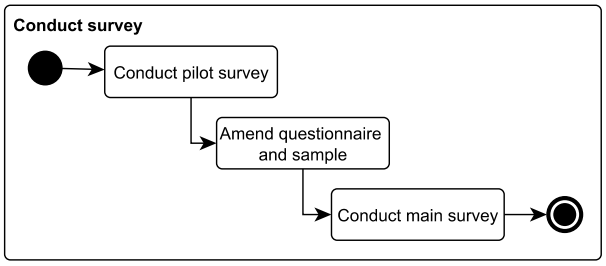


Figure 5.2: Activity diagram for *Conduct survey*. Adapted from Gray [73]

The piloting step is mainly intended to test the questionnaire's suitability. While we described the building blocks of the questionnaire in the previous section, our method gives the requirements engineers the flexibility to choose and combine them in the ways best suited for the measurement project. The feature descriptions are always project-specific.

For the testing of the questionnaire, a small number of testers (3-5 people) is asked to fill it in under realistic conditions. Their task is to note down anything they see as problematic with the questionnaire, such as ambiguous formulations. They are also asked to keep track of the time they need for each part of the questionnaire. Afterwards, the requirements engineers should have a short unstructured interview with the goal to find out how the testers felt about the questionnaire and whether they have major critique. The requirements engineers should pay attention to the scores chosen for *perceived understanding*, and discuss with the tester what formulation might lead to better understanding. This is standard procedure for piloting a study, as described in Gray [73]

In the next step, the requirements engineers amend the questionnaire. They use the information gathered during the test to change the formulations for better clarity, and to adjust the volume of the questionnaire such that the questionnaire can be filled in the allotted time.

The test and amendment step are then repeated. This iterative process continues, ideally until the requirements engineers feel satisfied with the questionnaire quality. There is no objective measure for this, as it depends on the requirements engineers' expertise. In practice, the available resources (access to testers, project deadlines) are likely to be the limiting factor on the number of iterations. When the questionnaire has produced good results, Gray suggests a last test with a large number of testers (20-40 participants). This step is not feasible in small surveys with access to a limited number of participants for the main survey.

The testers chosen for the pilot phase should be drawn from the final participant population. They should not be key users who have been actively involved in defining the requirements specification, or other members of the development team, because they are too well acquainted with the specification and their levels of understanding are not representative for the target population.

Gray suggests to also pay special attention to subgroups. If the questionnaire answers differ between testers representing different groups of users, the sampling frame for the main survey may be rethought such that sufficient users are represented from each of these groups.

Any other preparations (e.g. logistic concerns) should also be finalized during that phase. After that, the data gathering step can be carried out. Nothing in these steps is specific to our method, so we do not describe the details here. They can be conducted as suggested by generic survey literature, for example in [164].

5.3 Stage 3: Evaluate results

In the final phase of the survey, the requirements engineers evaluate the results and document them. The three steps needed for that are *prepare data*, *analyse data* and *write final report*, as represented in the activity diagram in figure 5.3. In this section, we describe the first two steps, preparing and analysing the data. Creating a report from the results is not specific to our method, and we do not describe it in detail.

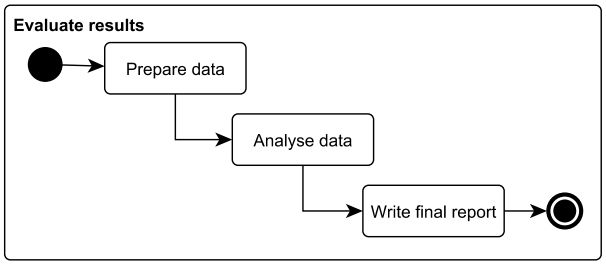


Figure 5.3: Activity diagram for *Evaluate results*. Adapted from Gray [73]

5.3.1 Prepare data

The information gathered through the questionnaire is at first available as crossed boxes on a printed answer scale, or an equivalent electronic representation. To make it suitable for evaluation, the requirements engineers should transcribe it in a single tabular data structure, as shown in table 5.3.

ParticipantID	Participant Group	Variable	SubjectID	Answer Numeric	AnswerString
1	novice user	usefulness	F1	1	1 - very useful
...	...	...	...	...	...

Table 5.3: Suggested format for the data collected from the survey.

Each answer is entered on its own row. The meaning of the columns is:

- The first column is a unique identifier of the participant. Names should be avoided in favor of pseudonymization for ethical reasons.
- This column is needed for later subgroup analysis. It represents a segment from the target population to which the participant belongs. If no subgroup analysis is desired, it can be skipped. Alternatively, there can be multiple such columns, if the requirements engineers plan to use different categorizations of the users.
- The second is the name of the variable being measured with the question, which we assume to be unique within the study. "Unique" in this context means that there are no two questions on the same feature that measure the same variable.
- The third is a unique identifier for the subject being measured, which can be a feature, a data record, or the system itself. The combination of a variable and a subject is sufficient to identify each instance in which a given question appears in the questionnaire. Although not technically necessary, we recommend that the subject ID encodes the type of the subject, e.g. by containing the letter F for feature, D for data and S for the system, as this makes the analysis easier.
- The column *AnswerString* has the name of the chosen answer option on the ordinal answer scale. We suggest that it is prefixed with the number of the option on the scale, because this allows the analysis software to automatically order the options correctly, and also because our scales do not use labels for all options. This value is the one primarily used in the analysis, since it reflects the correct level of measurement.
- For some of the summarizing results discussed in section 5.3.2, the answers are assumed to approximate interval data. For calculating the results, they have to be represented in a numeric format, not a textual one. Therefore, we recommend having a column with a numeric representation of the answer added to the table, which we call *AnswerNumeric* in the example.

From the example row shown in the table, we can learn that the participant with pseudonymous ID 1 answered a question which asked about the usefulness of a feature labelled F1 with the first option of the answer scale, corresponding to the label "very useful".

5.3.2 Analyse data

After the data has been collected and structured, the next step is to analyse it. The goal of this step is to answer the research questions, which are study-specific. In this section, we give a basic suggestion for interpreting the answers, and then show how this can be applied to aggregated data. The requirements engineers can then choose the aggregation criteria such that the evaluation is suitable for answering their research questions.

5.3.2.1 Interpreting the answers for a single questionnaire item

The most basic information that can be found from the data is the level of a given variable for a given subject. The formally correct summary statistic for ordinal data is the median, however with only five scale options it is not very informative. For a good understanding of the participants' opinion of the feature, we recommend visual inspection of the answer distribution.

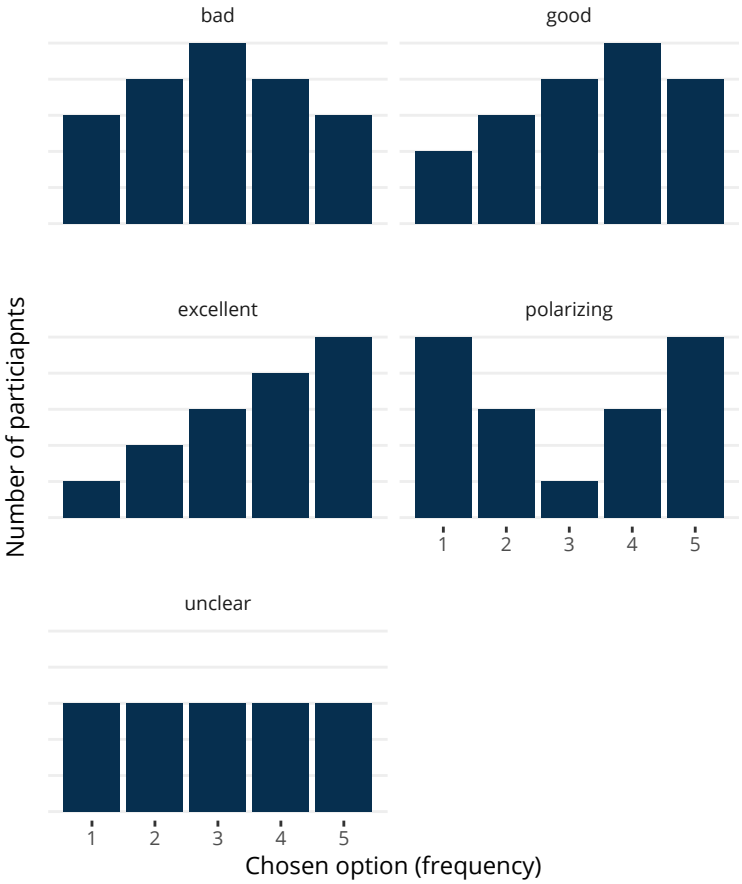


Figure 5.4: Evaluating a feature by answer distributions

In satisfaction measurement, the typical distribution is unimodal, with a peak in the positive side of the scale. Because of this strong positivity rating bias, a distribution with a mode around the neutral option describes a bad feature, a mode above the neutral option describes a good feature, and a monotonically increasing distribution with a mode in the highest option describes an excellent feature [153]. These three possibilities are depicted in the first three panels of figure 5.4. We also consider cases with a mode below the neutral position to be "bad".

In our studies, we sometimes see a U-shaped distribution, as shown in the fourth panel of figure 5.4. It has many answers on the positive and negative side, and less in the middle. We call this a *polarizing* feature. It shows that part of the users are very happy with the feature, while others are strongly against it. In this case, we suggest that the requirements engineers also look at the demographic data of the users who chose the best and worst options, and see if there is some difference between them. This can be an important tool for recognizing conflict within the user base.

If the distribution does not fit one of those patterns, for example by being closer to an equal distribution than any of the above (such as the last panel in figure 5.4), we cannot give more guidance for interpretation without knowing the context. It is possible that it is caused by a mixed form of the above cases, or that the participants are confused and either choosing answers at random, or choosing them in accordance with one of multiple possible interpretations. If an unusual distribution appears, it cannot be evaluated with the available results. We recommend reexamining the clarity of the question and discussing the values with that variable in person with key users, to better understand what motivates their attitude to it.

5.3.2.2 Numeric scores of satisfaction

There is a long-standing debate on whether it is permissible to calculate numeric scores for satisfaction variables. The usual method employed is to calculate the arithmetic mean of the answers to a given question. This calculation is widespread in practice [179], and is very popular with stakeholders, but there are strong arguments against using it.

The main appeal of using a numeric score is that the requirements engineers and the study's result recipients are well versed in working with numbers. The requirements engineers have the knowledge and the tools to represent them nicely. The study recipients can effortlessly perform mental operations such as comparing two scores, or follow the change in a score over time. The understanding and handling of a single number per variable only requires skills which are commonly taught in early school years, and it always provides an answer to some of the most common questions of result recipients, such as "Is feature A better than feature B".

In comparison, understanding and interpreting distributions on ordinal scales requires knowledge which is generally taught in tertiary education, and so the results recipients are less likely to be skilled at such interpretation. Also, the results exhibit less precision, and do not always

offer a clear-cut answer to some questions. If feature A and feature B are both seen as "good", it is difficult to determine if one is better than the other.

We find however that the desirable properties of numeric scores are misleading in the case of satisfaction measurements. First, they do not correspond to something that exists in the real world. The reason why satisfaction and its related concepts are measured with ordinal scales is that there is no physical property which can be measured in known units. The people responding to the questionnaire items think in terms of "I am very satisfied with this feature", not "I am 4.4 satisfied with this feature". When turning the ordinal measurements into a number, the meaning of that number is unclear.

A second problem is that an arithmetic mean calculation assumes at least interval data, that is, it assumes that the distance between each number is equal. There is no evidence that this is correct with satisfaction data. Instead, the different use of low and high scores suggests that the relationship is far from linear, which makes the arithmetic mean misleading – for example, a score of 3.5 does not mean that the feature's satisfaction is halfway between 3.0 and 4.0.

Third, the higher precision of a numeric score is spurious. There is evidence that study participants faced with 9 or more points on a scale have difficulty choosing an answer [153, 177]. The arithmetic mean of the answers of a 5-point scale suggests much higher precision, even if rounded to a single digit after the decimal point. So, if feature A has a score of 4.4 and feature B has a score of 4.2, the result recipients are likely to think that users are more satisfied with feature A than with B, while the users cannot tell the difference. Thus numeric scores can lead to false conclusions.

Because of these problems, we find that the use of numeric scores is misleading. The supposed easy understanding is likely to lead to overconfidence and wrong conclusions. In our work, we do not use them, and report the work on an ordinal scale.

5.3.2.3 Answering the research questions

Depending on the research question of the study, there are multiple options for the evaluation of the data. Here we present some approaches which correspond to the research questions suggested in section 5.1.2.

A/B test The first typical research question is to conduct an A/B test. In this case, there are only two features (actually feature variants) to be compared, and a high level of detail is desirable. Thus we suggest to represent the distribution of the answers for each variable for both features side-by-side. This representation, known as *small multiples* [178], allows good judgement when observing the variable levels for two subjects.

The comparison is done by comparing the distributions. If, in a given variable, the two features

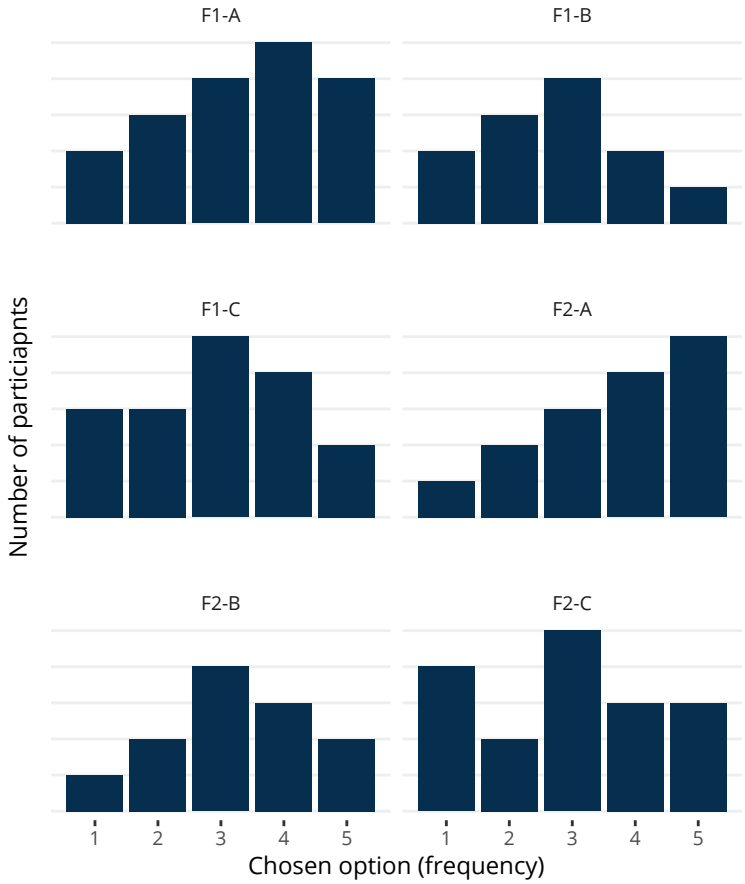


Figure 5.5: Comparing the results of two features, F1 and F2, for three variables A, B and C

have different modes, the feature with the mode at a higher option is better liked. If they have the same mode, then the distribution with the heavier right tail suggests better satisfaction.

This comparison method is illustrated in figure 5.5. It represents three fictional pairs of answer distributions for different features, assuming that positive answers are on the right side of the scale. In the examples in the image, for variable A, feature F1 is better than F2, because its mode is farther to the right. Distribution F2-A and F2-B both have the mode in the central position, but distribution F2-B has more answers to the right of the mode while F1-B has more answers to the left of the mode. This lets us conclude that feature 2 has better results for variable B. For the last pair, F3-A and F3-B, it is impossible to recognize a clear superiority of one feature over the other, so the comparison does not yield a conclusive result.

Problem magnitude of unpopular features This analysis is done for features suspected to be unpopular. It again uses the small multiples representation, but does not go to the detail level of comparing the distributions of each variable. Instead, an evaluation is created for each variable, as described in 5.3.2.1, and the results are plotted in one subgraph of the small multiples.

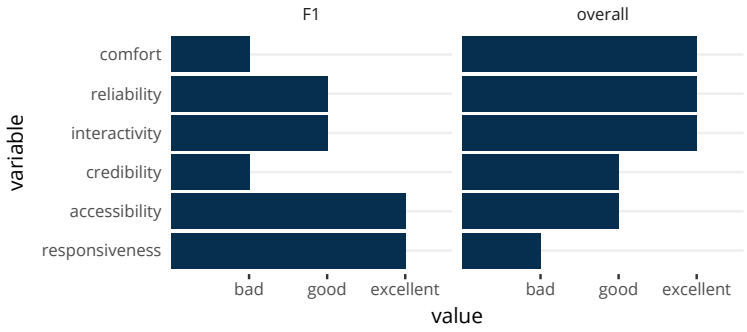


Figure 5.6: Analyzing the results for an unpopular feature F1

Then the procedure is repeated for the aggregated data over all subjects, for each of the variables measured for the analyzed feature.

An example is shown in figure 5.6. The feature F1 is low in comfort and credibility, and excellent in accessibility and responsiveness. This is very different from the overall pattern in the comparison chart. This difference is a sign that there is a real effect and not just a pattern typical for the dataset as a whole. On the whole, there are fewer good and excellent ratings of F1 than in the overall dataset, so the feature is indeed not as popular as others. The variables with low ratings give a good place to start searching for the cause of this issue, or for developing a solution to the problem.

In this example, there are no polarizing features. However, they are likely to occur with unpopular features, so we suggest being especially aware of them. Frequently the team may already have an intuition that a feature is unpopular with a particular role, for example with expert users. In this case, the analysis can be repeated for the participant subgroup defined by this role, using information from the preliminary questions to filter the answers.

If the feature is expected to be popular with some stakeholder groups, but not others, we suggest to inspect the answer distributions separated into these groups. This can be done on the level of individual variables, or aggregated across all variables. An illustration of this comparison is shown in figure 5.7. In this fictional example, the users like the feature more than the administrators.

Finding problematic features For finding problematic features, we use a two-step process. First, we have to identify the features with the worst distribution. To do this, we make a comparison similar to the one done for the A/B test, but using data which has been aggregated by feature, across all variables. This gives us one distribution per feature, and the height of each bar is equal to the sum of answers at that level for all variables.

These distributions can be then compared and the features with the worst distributions can be chosen for further consideration. The detail analysis is the same as the features preselected by the

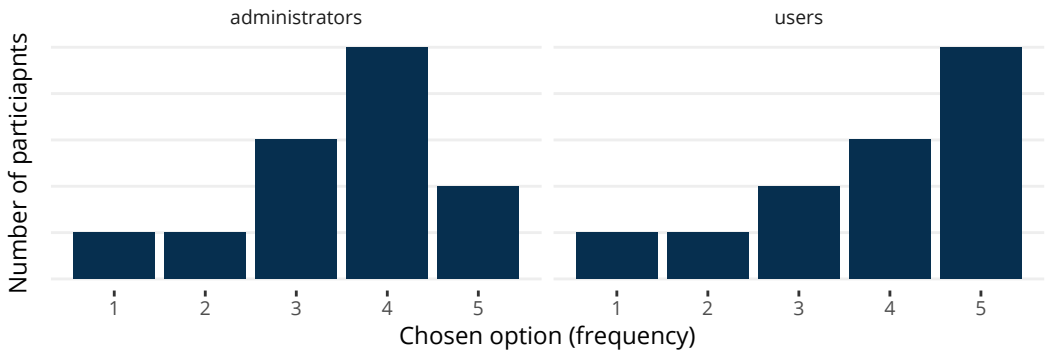


Figure 5.7: Comparing stakeholder subgroups

stakeholders in the *problem magnitude of unpopular features* section above. The features whose distribution has a heavy left tail (unpopular features) should be visualized in plots showing the answers to each variable, while the features with U-shaped distributions (polarizing features) should be shown in distributions dividing the answers by stakeholder group.

5.3.2.4 Additional evaluations

The analysis related to answering the research question is easier understood in light of the study context. For this, the requirements engineers should also include a very aggregated overview of the measured satisfaction levels and interesting findings from the open-ended questions.

The summary level is best visualized using a boxplot or a violin plot. The boxplot is a widespread plot type and stakeholders with intermediate knowledge of statistics will be able to read it without further explanations. Details on the structure and interpretation of boxplots can be found in [194].

Violin plots [88] are less common than boxplots, but add more information with a minimal increase in visual complexity. Figure 5.8 demonstrates a violin plot. It shows one symmetric shape ("violin") per feature. The "violin" can be interpreted as a smoothed density graph, rotated on its side and mirrored for better aesthetics. The Y-axis represents the 5-point answer scale from the questionnaire. In the example, feature F1 has a density function which starts flat at the low answers (the same number of participants chose an answer of 1, 2 or 3 on the scale), has some more answers at 4 and even more at 5, creating a funnel shape with a mode at 5, or a "very good" feature. Feature F2 has no answers at 1, few answers at 2 and 3, but a clear jump at 4 and very few again at 5. It is widest at the level of 4, so this is a "good" feature. The shorter shape with a wider mode shows that there is less variation in the answers than in F1. For comparison, the second panel in figure 5.8 shows the same data as a boxplot.

An important information source in the study are the answers to the open-ended questions. They can give new insights for the development team. Beside just using the full text as an opportunity

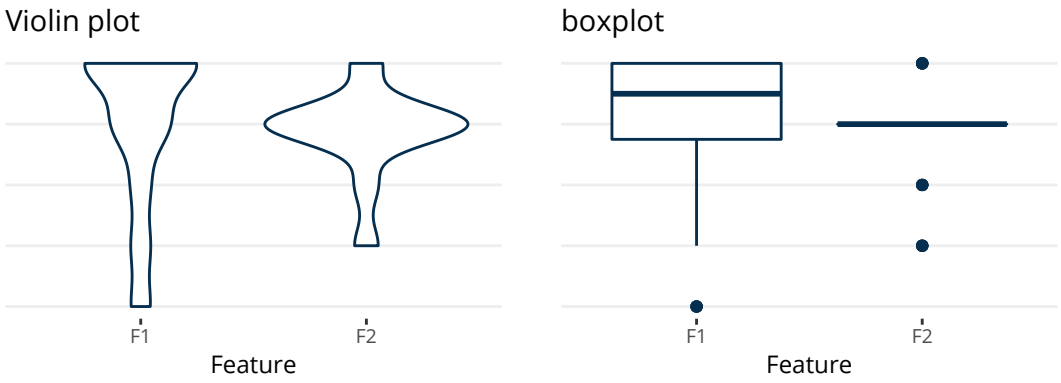


Figure 5.8: A boxplot and a violin plot for aggregated results, both showing the same data

to learn from the users, we also suggest that the requirements engineers also decide on the information needed to answer their research questions, and code the free-text answers for that information. For example, if the development team considers the possibility of data export in a given format but does not know how needed this is for users, the requirements engineers can include a description of the export feature in the questionnaire, add a field for free-text suggestions to it, and then code all answers as either requesting that format or not. Then the frequency of requests can be presented as a result answering the study's question.

5.4 Summary

In this chapter, we describe a method for measuring anticipated satisfaction. It is a survey-based method, consisting of designing questionnaires, distributing them to users, and evaluating the users' answers.

The questionnaire contains descriptions of the system's purpose, the data processed with the system, and its features. The questions in it measure both discrepancy variables such as *perceived understanding* and variables derived from satisfaction concepts such as *usefulness* and *credibility*. The discrepancy variables are based on the model of anticipated satisfaction we develop in chapter 3. The satisfaction variables are chosen from a list of concepts used in the satisfaction literature. We provide such a list in chapter 4.

We also give guidelines for analyzing the data collected with the questionnaire. The answers for each questionnaire item can be categorized on a simple *good-bad-excellent* scale, which best reflects the precision of the information and is also easily understandable for result recipients. For answering the research questions, the data is usually aggregated in appropriate ways before evaluating the result. The free-text data can be reported as-is, to serve as salient examples of the users' opinion, or coded as needed for the research questions. The final results are then summed up in a report.

6 Validation of the method for measuring anticipated satisfaction

In the last chapter, we presented a method for measuring anticipated satisfaction. We intend it to be used as a validation approach for software requirements. The goal of this chapter is to empirically evaluate the method's usability when applied in a software development project, thus showing that it is suitable for its purpose.

We base our validation on a list of criteria for evaluating usability methods proposed by Hartson, Andre and Williges [80]. This is an ideal list, and even its authors suggest that it is difficult to measure some of the criteria in reality. We do not cover it in full, but choose to use the four criteria of *validity*, *downstream utility*, *cost effectiveness* and *usability*.

Our validation approach consists of applying the method in two empirical studies. In the first we apply the method ourselves, to show that the prodcedure is feasible and acceptable by the users. This study also includes a measurement of actual satisfaction, which allows us to calculate how well anticipated satisfaction reflects actual satisfaction. A second method application involving a professional outside of our research team additionally shows the method's learnability and its potential for industry acceptance.

The chapter has four sections. The first gives an overview of Hartson's criteria. Sections two and three describe the two studies we conducted to validate our method. In the last section, we discuss the threats to validity and summarize our conclusions about the feasability of applying the method in development projects.

6.1 Planning the evaluation

Since a satisfaction measurement method is a special kind of a usability evaluation method (UEM), we can apply UEM evaluation criteria to validate our method. We base our validation on the criteria proposed by Hartson, Andre and Williges [80].

Reliability is present when a UEM gives consistent results. This is a basic quality requirement for all measurement methods, but can be difficult to achieve due to differences in measurement context.

Thoroughness means that an UEM can "find as many of the existing usability problems as possible". It is an attractive criterion, but difficult to measure in practice, as the real number of usability problems is rarely known.

Validity is another metric desired of all measurement, not only usability. In a UEM, it means that the UEM only finds real problems and does not produce false positives.

Effectiveness Hartson at al. suggest a compound quality metric, $thoroughness \times validity$, and name it *effectiveness*.

Downstream utility UEMs are used to provide knowledge relevant for improving the system. Their contribution to that goal can be elicited from the project members who applied them, and from the system's developers.

Cost effectiveness Cost is an important consideration in the business world, so it should be considered when choosing an UEM.

Usabilitiy Hartson and his colleagues have found no instance of a UEM's usability being measured, but suggest that it should be a natural thing for usability researchers to measure about their own tools.

There is no practical way to measure all of these criteria. Instead, we chose to measure the more accessible ones – validity, downstream utility, effort and usability.

The criteria of downstream utility, effort and usability are best measured if a study follows the guidelines of the last chapter exactly. For reliable results, the context of the study should be representative of a real method application – there should be a real development project, the roles described in section 5.1.1 should be filled by the real stakeholders and not stand-ins, and the steps of the method should be followed as prescribed in the guidelines.

This presents a problem for measuring validity. Our method produces a measurement of anticipated satisfaction. To demonstrate validity, this measurement has to be compared to a measurement of actual satisfaction. However, such a measurement does not exist in a system which is still under development.

The most straightforward possibility would be to apply the method during development, then wait for the system to be finished and for the product owner to conduct a usability study which produces data on actual satisfaction. This was not possible for logistical reasons, specifically the constraints of the duration of a PhD project and the limited availability of development teams interested in cooperating in such a study. It would also introduce methodological concerns, especially the question whether the data from a separate study, presumably elicited from different participants, is comparable with the data from the first study on anticipated satisfaction.

Instead, we chose to conduct two separate studies. The first one, which we call the Casino study,

is a two-part study suitable to measure validity. It started when the system was in the final stages of development, and gathered first data on anticipated satisfaction, and then, after the system was released, also gathered data on actual satisfaction. This allowed us to demonstrate the validity of anticipated satisfaction measurements.

Since the Casino study included activities for the measurement of both anticipated and actual satisfaction, it could not produce an unbiased estimate for effort. Only the pre-exposure measurement would be relevant for comparison with a usual project based on the MUSA method, but most activities in the Casino study could not be clearly divided into belonging to one or the other measurement.

There was also no person connected to the project who could assume the requirements engineer role, and the thesis author had to fill the role. This undermines our findings on the usability criterion, since the ways in which the thesis author relates to her method are not representative of other requirements engineers employing the method. It was also impossible to measure the usability for the participants, since they were anonymous in this study.

In our second study, called the MITO study, a member of the development team assumed the role of the requirements engineer in most stages of the method. When conducting the study, we strictly followed the guidelines for MUSA and did not gather additional information. This allowed us to gather better quality evidence for effort and usability. We also measured the downstream utility, and so have good evidence for it from both studies. Since there was no post-exposure measurement of actual satisfaction, we were not able to measure validity.

The two studies complemented each other, allowing us to cover all four of our chosen criteria. With those criteria, we can address major concerns both from a theoretical perspective (does the method deliver what it intends to) and those arising in practice (is it feasible to apply the method within the logistic constraints of a typical development project). This makes our evaluation strategy appropriate for a novel software engineering method.

6.2 Casino study

In this study, we applied our method on a new system shortly before its first release, then used a traditional questionnaire to elicit actual satisfaction after the release. This allowed us to apply two of the criteria for evaluating usability methods, *validity* and *downstream utility*. It also gave us some basic evidence on two further criteria, *usability* and *cost efficiency*. The concept was similar to the studies described in chapter 3, with the difference that this time we applied the completed method instead of a collection of candidate questions, and the actual satisfaction was measured after real system use as opposed to viewing a tutorial.

We formulated a research question for each of the four criteria. The questions conform to the definition of their criterion, but are more concrete to reflect the specifics of our study context.

RQ8: Validity How well do our measurements of *anticipated satisfaction* correspond to the measurements of *actual satisfaction*?

RQ9: Downstream utility How can the results of our method be used in the development project?

RQ10: Usability Were there major usability problems during the method application?

RQ11: Effort How much effort was needed for conducting the study?

To show *validity*, we have to compare the measurements of our new method to measurements obtained with an established method. In this case, we use a standard method for *actual satisfaction* as the standard of comparison, and report the deviations between the two measurements in section 6.2.2.1.

The downstream utility is best judged by the stakeholders. We interviewed the team members and the product owner of the system as potential beneficiaries, using an open-ended technique. The results are summarized in section 6.2.2.2.

In the same interview, we also inquired about the *usability* of the method and combined the information with our own observations collected when conducting the study. A summary is given in section 6.2.2.3.

We measured the *effort* needed to execute the study in worktime, since this is likely to be a scarce resource in a development project. In section 6.2.2.4, we give a description of the time spent by all affected roles in the different steps of the project.

6.2.1 Materials and methods

The DKFZ (German Cancer Research Center – Deutsches Krebsforschungszentrum) is a research institute with 3600 employees. Its canteen (officially named the Casino) offers a catering service for academic events organized by its departments. In 2016, it replaced the paper-based process for catering orders with a bespoke online system developed in-house.

The DKFZ Catering application automates a well-defined process. The customer (who is usually versed in organizing events for their department) opens an ordering form and enters the basic data for the event, such as date and place, and the desired amount of food and drinks. The canteen confirms the order and updates the status in the application, then prepares the food on the desired date. The application shows the price for information purposes, but does not implement the billing process.

6.2.1.1 Process

We started by preparing two questionnaires. The first one was the pre-exposure questionnaire, which measured anticipated satisfaction and corresponded to the single questionnaire described in chapter 5. The second one, measured post-exposure, was used to determine the actual satisfaction after use, which allowed us to measure the *validity* of our method. It also served as an additional incentive for the product owner to agree to the study. Since the system was almost complete at the time the study started, he was more interested in the measurement of actual satisfaction than of anticipated satisfaction.

Both questionnaires were prepared by the author, who followed MUSA for the first questionnaire. The second questionnaire was prepared by the author, closely following the content of the first to allow comparability.

The use of a newly developed application ensured that none of our participants had any experience with it when filling the pre-exposure questionnaire. It also let us address the entire population of potential participants, by advertising the study in the internal mailing list. This advertisement invited the participants to fill the pre-exposure questionnaire. In this questionnaire, they were asked to identify themselves with a pseudonym derived by a repeatable algorithm, but not reverse-engineerable by other stakeholders of the project.

After the system was released, the pre-exposure questionnaire was deactivated. When a user completed an ordering process with the system, the system waited until after the delivery date of the order, then sent an invitation to fill the post-exposure questionnaire. This ensured that the participants who filled the second questionnaire had indeed used the system.

After the data gathering was complete, the author evaluated the study according to the ASMA guidelines. She discussed the results of the first questionnaire with the development team to determine their impressions of the method use. Beside the results for this thesis, she also produced a report on the actual satisfaction with the system, to be used by the product owner.

6.2.1.2 Stakeholders

The study had the usual stakeholders described in section 5.1.1. The author conducted the study, filling the role of the requirements engineer. The study participants were drawn from the system's target population, although many of them were novices who did not know the process well. The test users were friends of the author, not acquainted with the system.

The whole development team consisted of one developer and one project manager, and both were involved in the study. The product owner, who is head of the department which supplies the food ordered through the system, and the DKFZ management board were the result recipients. The product owner also played the role of champion together with the project manager. Finally,

	activity	stable properties	mutable properties	appraisal
user	– use	– personal innovativeness	– need fulfillment	
system	– effectiveness	– complexity		– system rating – helpfulness
information		– understandability	– relevance – timeliness*	(not used in study)
context		– alternatives	– benefit* – task outcome	

Table 6.1: Concepts used in the DKFZ Casino study. Concepts marked with an asterisk were requested by the product owner.

the data protection officer and the *staff council* [1] were revisors of the study, since the author had to show to them that the study conforms to data protection and employee protection regulations before being allowed to contact the participants.

6.2.1.3 Questionnaire

We identified one data record and two features to evaluate in the first questionnaire. The first feature allows customers to request food or drinks which are not listed on the standard menu. This is a rarely used feature and the recipients were interested in knowing if it is really needed. The second feature was that the application sends status change confirmations by email. The product owner had reports that some users feel spammed by the mails and others feel reassured by them. He was interested in gathering more data about the situation.

The data record we chose was the ordering form for food and drinks. Since it is the focal point of the system, it has a high impact on how the users interact with it. Also, it would have been difficult to introduce the concept of special wishes in the description if the standard menu items (which are part of the ordering form) are not presented in the questionnaire.

The study did not have a specific focus on given variables, thus we used the default option to cover all ASMA categories. This produces a broad view on the system, allowing us to elicit information about different aspects. The one exception was that we did not use the *information appraisal* category.

The product owner approved our choice of variables. For him, they all reflected an interesting information about the system. He agreed that *information appraisal* is not very relevant for the

[1] German legal term: *Betriebsrat*, an internal committee for upholding employees' rights

system goals and does not have to be covered. He also had a special interest in two additional variables, *benefit* and *timeliness*, and asked us to add questions about them to the study. The final variable selection is shown in table 6.1, and the full questionnaire is reproduced in The full questionnaire is reproduced in appendix B.

The post-exposure questionnaire used items for the same variables, plus the new variable *disconfirmation*, which is calculated as the difference between the users' expectations and their actual feelings about the system.

6.2.2 Results

We received responses from 47 participants. Of those, only 6 participants filled both the pre-exposure and the post-exposure questionnaires. There were 33 more responses only to the pre-exposure questionnaire and 8 responses only to the post-exposure questionnaire. We had indications that this discrepancy is due to the different channels in which the study was announced.

6.2.2.1 Validity (RQ 8)

To answer RQ 8, we made two separate evaluations. First, we concentrated on the responses of participants who have answered both questionnaires. As this only covered 6 participants, we could not run a parametric hypothesis test. Instead, we chose to use a descriptive method.

The questionnaires were pseudonymized, which allowed us to match the answers of the pre- and post-exposure questionnaire for each participant. That allowed us to compare each person's change in opinion between the two questionnaires, instead of comparing only answer averages, increasing the power of our analysis.

We calculated the difference in the pre- and post-exposure evaluation of each item for each participant. We left out the two user-related items *use* and *personal innovation*, since they do not depend on system exposure.

The differences could range between -4 (user chose a 5 in the pre-exposure and a 1 in the post-exposure questionnaire) and 4. In our data, we only observed differences in the range between -2 and 2. 56% of the answers showed no difference, and 30% had a differnce of -1. The Spearman's correlation between pre-exposure and post-exposure answers was 0.71.

These numbers show high agreement between anticipated and actual satisfaction. If there is a difference, it is usually a single point reduction. This hints that users have a mostly image of the system before use, with a trend towards idealizing it.

6.2.2.2 Downstream utility (RQ 9)

The study was well received by the result recipients. The product owner found the detailed report easy to understand, despite having no formal knowledge about statistics. The project manager was slightly concerned about the actual satisfaction being worse than expectations, but in a talk with the product owner concluded that this is not a sign of project failure, since users frequently have overly optimistic desires for new systems. All results recipients found the information interesting, and after a lively discussion, the team generated new requirements to be implemented in the next version of the system.

We interpret the product owner's active participation in the questionnaire design as a positive sign. He found our chosen variables informative, which shows that full ASMA coverage is a viable option from the recipients' point of view. The results for the variables in which he was personally interested also fulfilled his expectations. He was equally positive about our choice of data record and first feature, and suggested the second feature himself.

The interest in both the preparation and the results show that the result recipients find this type of study useful, and that it has utility for their work as managers. Similarly, it has utility for the development team, which can decide on features and their detailed implementation in a much more directed way, compensating for weaknesses seen during the study.

6.2.2.3 Usability (RQ 10)

To follow our method, we had to manage the satisfaction measurement project, create a questionnaire, gather data, evaluate it, and communicate the results to the recipients. The project management part was not difficult, as the project found good acceptance among the development team and the product owner. Designing the questionnaire was also not difficult.

We did not gather data on the usability of the data gathering step from the participants, since they were anonymous and we could not contact them for comments.

The result evaluation was less straightforward than the questionnaire creation. As there is no standard software for creating the evaluation, we had to program a custom solution for evaluating the scripts.

Communication with the product owner and the project manager was also not completely straightforward. It was predominantly positive, but some explanation was needed for them to understand the result format.

6.2.2.4 Effort (RQ 11)

The effort for the study varied between the different steps. The questionnaire creation took the least effort, about two workdays for the requirements engineer, plus a few hours for coordinating the content with the stakeholders. The questionnaire was created directly in LimeSurvey and the effort for typing and layouting is included in the effort for creating it. Since the software allows the participants to access the software on their own and has a built-in export functionality for the answers, the data gathering step did not require any effort from the requirements engineer and the development team.

Evaluating the answers was the most time-consuming part of the process. It took about two workweeks to prepare the data and create all necessary calculations, graphs and a report. [2]

The project management and communication with the recipients was done throughout the project, and required several one-hour meetings. While it did not require many workhours in total, it needed long waiting times due to the complexity of coordinating decisions between multiple stakeholders.

6.3 MITO study

In this second study, we applied our method to a software system for biology researchers. We used the same criteria as for the Casino study, but changed the study design as described in section 6.1 to get better evidence for effort and usability, and did not measure validity since this was not possible with the new design.

RQ12: Downstream utility How can the results of our method be used in the development project?

RQ13: Usability Were there major usability problems during the method application?

RQ14: Effort How much effort was needed for conducting the study?

Similarly to the Casino study, we give a qualitative description of the usability and downstream utility of the study based on an interview of the stakeholders, and we measure the effort by the time needed for the study.

[2] This time refers to creating a report as prescribed by MUSA, which is reproduced in the appendix. For the validation, we did two more evaluations. One was the qualitative and quantitative evaluation described in this section, and the other was a report on the postexposure data created on request of the product owner. These are not part of an application of MUSA and are not included in the time reported here.

6.3.1 Materials and methods

The system under test in this study was MITO, a specialized database for biologists. It facilitates the search for biological model systems [3]. It was developed in-house at the DKFZ, and the thesis author was part of the development team.

The MITO system stores three main data records. One of them is the *tumor model*, which describes a protocol for creating a given model system. For example, it can include transplanting cancer cells into a mouse, waiting until the mouse has a tumor, then treating the tumor with a drug candidate. This kind of protocol is only successful if the same strain of animal and/or the same cell line is used every time. Thus, MITO also contains the data record *animal line* and the data record *cell line*, which describe the organisms used in the tumor models.

MITO's target users are researchers who want to try new interventions in an existing tumor model. They use MITO to search for existing models and possibly compare them before choosing one for their work. This makes the search for tumor models, animal lines and cell lines the most used functionality in MITO. Its further functionalities are related to data entry and managing users and read/write permissions.

6.3.1.1 Stakeholders

For this study, we asked a member of the development team to assume the role of the requirements engineer. His usual position on the team is to be a software developer, but as this is a small team, he has also been involved in eliciting and updating the requirements for the system. He applied our method with assistance from the thesis author. Project management and evaluation were done by the thesis author, while the requirements engineer created the questionnaire, communicated the results to other stakeholders, and used the results in his own work on the development project.

The participants were biologists in their fourth week of a master's programme at the DKFZ. As they had only recently arrived at the DKFZ, they had not yet used the system, and thus could fill a questionnaire of anticipated satisfaction. The study was championed by MITO's product owner, who allowed the author to gather answers and interview the participants as a part of a lecture conducted by the product owner.

The student status of the participants means that they are not completely representative of the target population. They have not yet started designing their own tumor models, so they have not yet needed to search for existing ones in their everyday work. Nevertheless, their background is sufficient to know what a tumor model is and how it is used, and to understand its description.

[3] This meaning of *system* is unrelated to the term *software system*. It typically refers to organisms used for biological research

The results recipients consisted of the product owner and the development team including the project manager.

6.3.1.2 Process

For this study, the requirements engineer was given the guidelines for MUSA and asked to follow them. He used them to prepare a questionnaire for the participants.

The data gathering step was done offline. The participants were attending a lecture on biological model systems followed by an exercise for which they had to use the MITO system. After the exercise, the thesis author asked them to fill the questionnaire. After collecting the questionnaires, she interviewed each participant individually on the usability of the method from their point of view.

The thesis author transcribed and evaluated the data. The results were discussed with the development team and the product owner, who used them to derive insights about possible changes to the system. There was no written report, since the result recipients indicated that they do not need it.

6.3.1.3 Questionnaire

The questionnaire contained questions on all three data records - tumor models, animal lines and cell lines. It covered two features, *general search* and *advanced search*. The requirements engineer chose those two features since the system is mainly used to look up recorded information, so search is a core functionality and critical to system success. Together with the introductory questions and the questions on the system as a whole, it contained a total of 38 questions. The requirements engineer decided against including further features in the questionnaire due to length considerations.

The requirements engineer used the ASMA variables to create questions, but also added some new variables of personal interest, for example *standardization* (a question asked if the users prefer the data records to use a standardized format). He also used many variables more than once, for example *content*, asking detailed questions about different parts of the data record. He did not attempt a full coverage of the ASMA categories.

A list of the covered variables is shown in table 6.2. The full questionnaire is reproduced in the appendix.

	activity	stable properties	mutable properties	appraisal
user	– use	– personal innovativeness	– attitude	
system	– efficiency	– preference		– helpfulness
information		– completeness – content – media richness – standardization – understandability	– (not used in study)	(not used in study)
context		– alternatives	– benefit – task outcome	

Table 6.2: Concepts used in the MITO study

6.3.2 Results

There were 9 participants, which was the full size of that year's class. All of them agreed to answer the questionnaire and to answer the subsequent interview on usability.

6.3.2.1 Downstream utility (RQ12)

The results of the study were presented internally and discussed by the development team. The overall anticipated satisfaction with the system was good to very good, and the team saw this as a positive sign. There was however concern that the results are biased due to the study population (graduate students) not being representative of the target audience for the system (professional researchers).

Some questions in the questionaire were specifically written with the purpose of discovering potential shortcomings in the system. For example, there were questions asking if the information given on a search results page is sufficient to decide which search result is interesting enough to be checked in depth. The team had been prepared to take action if these questions discovered problems in the system. However, the respondents did not indicate such problems, so no action was taken. For the team, this was a further sign that the system meets the users' expectations.

6.3.2.2 Usability (RQ 13)

We asked the requirements engineer to share his impressions in an unstructured interview. He found the guidelines easy to apply, and did not encounter any problems in creating the question-

naire. He liked the "flexible" format of the questionnaire, which allowed him to add questions about issues he had anticipated during development. An issue he encountered was that he was uncertain which variables from the ASMA table would be a best fit for his situation. Also, he imagined the evaluation to be complicated, and would have preferred a tool which automatically creates the evaluation without the need to learn about statistical tests. Overall, he had a positive impression of the process and stated he would apply it again in further projects.

We also measured the usability for the participants in a structured interview. After filling the questionnaire, they first watched a presentation of the completed software as a group activity, then they were asked the four questions below.

Was filling the questionnaire difficult or easy for you? Four participants reported that the questionnaire was easy. Only one stated that it was difficult. The remaining four said that it had both easy and difficult aspects.

Did the system as seen in the presentation match what you imagined when reading the questionnaire? Five participants stated no differences between the way they imagined the system and what they saw in the presentation. One pointed out that a certain functionality was very surprising to him. The remaining three felt there was some difference between the imagined and real system, but did not explain in detail how it differed.

Would you consider participating in this type of test again, if asked to? All nine participants gave a positive answer.

What was most difficult for you when answering the questionnaire? Eight out of the nine participants listed some points of difficulty here, with one stating that she did not perceive anything as especially difficult. The others usually listed multiple issues.

Two of the issues they listed were problems with the system itself and not with the usability method, for example one participant stated that he did not know what an ontogenetic stage is (this was a field in one of the data records presented). One was about the questionnaire design (did not always know which question referred to which field). Two participants observed that it was their lack of domain knowledge which made it difficult to judge whether the system is fit for the task described. Four more statements were an expression of feeling insecurity when trying to answer.

There were three issues which we saw as caused by our method itself. One was that a participant would have liked to add more information in some places. This is normal for all questionnaires which employ a scale instead of open questions, and can be offset with the inclusion of freetext fields. The second was that a participant felt an urge to give "nice" answers. This is evidence

that our method is subject to the effects of social desirability bias. The third issue was by a student who at first did not understand that she is supposed to imagine the system based on the description, and was confused how she is supposed to come up with answers.

Overall, we find that the method has good usability for both the requirements engineers and the participants. The most important issue is that the instruction to base their answers on their imagination might confuse the participants, and frequently leads to feelings of insecurity. Nevertheless, they report high willingness to participate in similar studies.

6.3.2.3 Effort (RQ 14)

The programmer reported that a workday was sufficient to read the guidelines and create the questionnaire using a specialized questionnaire tool. Transcribing the data only required a few hours for all nine participants (the questionnaire had 38 questions). The statistical evaluation needed three workdays, but it was done by the author, who was already well acquainted with the methods needed and could partly reuse existing R scripts. Creating a final report needed two more workdays.

It should be noted that only the data transcription effort scales linearly with an increasing participant number. The time for creating the questionnaire, evaluating the data and writing the report did not depend on the number of recorded answers, so it should stay similar for much larger participant numbers.

6.4 Discussion

In this chapter, we conducted two empirical studies in which our method was applied. This allowed us to investigate how well our method performs under realistic conditions, and to discover potential issues in its use. The first study consisted of us applying our own method, then measuring actual satisfaction with the same system and comparing the results. This allowed us to gauge the difference between actual and anticipated satisfaction. In the second study, a software engineer without previous experience with the method constructed a questionnaire for a system he was developing. This showed that the method can be learned by software engineering professionals without major difficulties.

6.4.1 Threats to validity

The work in this chapter is subject to limitations. We aspired to keep the study context as close as possible to the context intended for our method. Still, there were some differences which might have biased the results. Also, conducting the method only two times cannot give an in-depth

insight into all potential issues which might arise in the future. Thus the validation should be understood as a proof-of-concept that the method can be used with a positive outcome.

Construct validity The construct validity for both our studies is good. We measured the three variables of *downstream utility*, *usability* and *effort*. Straightforward measures exist for all three of them. For downstream utility, we recorded the reactions of the development team, and their intention to adjust the system specification based on insights from the study. Our usability questions covered standard variables. Effort was measured in workdays, broken down by role.

Internal validity We did not use validated instruments for measuring the three variables that answer our research questions. Rather, we used a qualitative approach which allowed us to capture the nuances of meaning as perceived by the people involved in the studies. We did not employ any quantitative evaluation which would have necessitated standardized data gathering methods.

External validity Both studies had elements which would not be present in a typical measurement of anticipated satisfaction In the first study, the questionnaire was prepared by the author, so it does not allow conclusions about the viability of the method when applied by an average requirements engineer. Also, the course of the study (which measured both anticipated and actual satisfaction) makes effort measures inaccurate, since the study contain more steps than the method foresees.

The second study suffered from a discrepancy between the target population (biology researchers) and the participant population (graduate students in biology). The results are obviously skewed by that, to the point where it was a topic raised by both the result recipients and the participants themselves. However, it still allows for a good estimate of the effort, and the effect on usability is likely to be an underestimation (students had more difficulty imagining the system and therefore found the questions more difficult to answer) than an overestimation. A further limitation of the second study is that the evaluation was done by the thesis author and not the requirements engineer. While it is possible that in a routine application of the method there will be multiple people in each role, the thesis author is not as representative of the requirements engineers as a person without connections to the method creation.

For both studies, we used systems which were very late in their development cycle. The preexposure questionnaire for the Casino system was disseminated a few days before the release date, while the MITO system was already released, but the participants had not had access to it. Typically, at that stage the development team already has had extensive feedback about the system from alpha testers and other sources. If the measurement is conducted at an earlier stage, it is possible that the requirements engineers would find it more difficult to choose good concepts to measure (because they do not yet know the most problematic points of their system), need more effort to create the questionnaire (because they do not yet have implemented features to

describe) and react differently to the study conclusions (since it is easier to change a system in the earlier stages of development).

Conclusion validity For most of our research questions, we did not conduct a quantitative analysis or hypothesis testing, so we cannot apply the standard criteria for conclusion validity (e.g. the magnitude of correlation coefficients or other numeric indicators). Rather, our study has descriptive results, from which we derive statements about the method's suitability for use in development projects.

The conclusion validity of our qualitative analysis is fair. The data gathered is not sufficient to exclude all potential alternative explanations for our conclusions. However, multiple details of our study support each other to make our conclusions a very likely interpretation of the observed data.

We had one question which was answered by quantitative analysis. This was RQ 8, which concerned the discrepancy between actual and anticipated satisfaction. For this question, the major issue was the low number of participants, with 33 participants in the preexposure questionnaire, 14 in the postexposure, and only 6 of those doing both questionnaires. However, we were able to compare the answers over a large number of questions, creating many data points despite the low number of participants. Also, we were able to do a paired calculation for the six participants who answered both questionnaires, resulting in a more powerful analysis. While more research is needed to investigate the exact nature of the connection between the two concepts, our study is sufficient to recognize the general trends.

6.4.2 Conclusions

We used a list of evaluation criteria for usability methods from the literature [80] and chose four which can be realistically measured and have an impact on the adoption of the method in a commercial environment. These are the *validity*, *downstream utility*, *usability* and *effort* as a measure of *cost effectiveness*, since our method does not require financial investments or other costs. Since our method measures a type of forecast, we interpreted validity as forecast power and compared our method's result – *anticipated satisfaction* – with the measurement it tries to forecast, the *actual satisfaction* with the same system.

The forecast power of the method is moderate. We discovered a discrepancy of about 1 point between measurements of anticipated and actual satisfaction, with anticipated satisfaction being typically higher. This is a substantial difference in a 5-point scale, especially in satisfaction measurement where answers typically only concentrate on the positive half [153]. This limits the use of our method for purely forecasting purposes, since the actual satisfaction is likely to differ after the system is built.

This reduced forecasting power does not make our method obsolete, since its main objective is to discover usability issues which can be corrected early on. We recommend to carefully communicate this phenomenon to the result recipients, to prevent inflated expectations.

Our explanation for this phenomenon is that users cannot foresee all potential issues with a system in detail, and requirements engineers cannot describe them all in the questionnaire. A feature can look good on paper, but it is always possible that an unforeseen factor will reduce its benefits for the user. This means that our method can never achieve perfect *thoroughness* (which is yet another criterion for a usability method). The teams employing the method should be aware of this limitation, and be prepared to employ additional methods in later stages of development to find the issues which were not discovered by our method. This does not detract from the usefulness of our method, since early discovery of issues is highly desirable.

This is also reflected in our method's *downstream utility*. The development teams in both studies showed high appreciation for the results. From their point of view, conducting the study had been useful. The Casino system could also be changed based on the study results. No formal metrics are available to measure the impact of these changes, but the development team and product owner voiced satisfaction with them. The MITO study did not change the requirements specification for the system, a decision which the development team contributed to the lack of domain expertise in the participants. This is a factor which should not arise in a typical application of our method, where we expect the participants to be drawn from the target population.

The next criterion was *usability*. In the first study, we did not conduct a detailed analysis of method usability, since the method was applied by the author, whose evaluation would be strongly biased. We merely noted that there were no obvious usability issues rising up. In the second study, we conducted interviews to determine the usability for both the person who had the requirements engineer role and for the participants. Again, no major issues were observed. The requirements engineer had a positive impression of the method's usability, merely mentioning some of the sources of complexity in the method, e.g. the need to choose which concepts to study. The participants also gave a predominantly positive evaluation. As expected, the most unusual part of the method – asking opinions on the basis of something imagined – was perceived as more complex and a source of uncertainty. Nevertheless, the participants unanimously declared that they would repeat a study of this type, and nobody abandoned the questionnaire due to usability problems. We conclude that the cognitive complexity of filling the questionnaire does not act as a barrier to study participation.

Lastly, we analyzed the method's costs, represented by *effort*. In both cases, the work needed was less than two person-weeks. We think that this is a good figure to use for planning for a measurement project. Due to a learning curve, the first application of the method within a given organization will probably take more effort than this. However, with growing process maturity, it is likely that this time can even be reduced.

Beside measuring these criteria, we also noted the overall attitude of the stakeholders to the method. It was predominantly positive, and they appreciated the flexibility of the method. The

choice of variables was noted as sometimes difficult, since it is not clear which ones fit best a given system.

In summary, our studies show that the method can gain acceptance in a commercial setting. It delivers results with reasonable accuracy, and its result recipients understand and value the information they gain by applying it. They tend to support the measurement project, and even suggest their own additions depending on the specific details of the project. The method is flexible enough to allow that, which increases the result recipients' satisfaction and their benefit from the measurement. The method also has good acceptance among the requirements engineers who create and evaluate the questionnaire and among the participants themselves. The unusual mode of answering does not deter them from participating. The effort for applying it is not trivial, but also not prohibitive when compared to the overall effort needed for a development project, and to the potential savings from reduction in late-stage corrections to the system.

Among the weaknesses of the method are the complexity of the evaluation, which requires intermediate knowledge in statistics, and the low thoroughness. It cannot be guaranteed that all issues with a feature will be discovered. However, this is information which probably cannot be obtained during requirements validation by any method, since the final satisfaction with a feature will be dependent on the concrete implementation. Thus we see our method as a complement to standard usability studies done after software release, and not a replacement for them. Its strength lies in the early validation, and this makes it a valuable tool in software development.

Part III

Summary

7 Conclusion and Future Work

The aim of this thesis is to create a method for predicting user satisfaction with a software system before the system has been implemented. We call this a measurement of *anticipated satisfaction*, to distinguish it from the *actual satisfaction* measured after exposure to the system.

The first result of our research is a list of satisfaction-related concepts. Using a systematic literature review, we compiled a list of 80 concepts which are used in satisfaction measurements, and conducted a metaanalysis for the 20 concepts for which we had at least 5 distinct sources.

We also created ASMA, a categorization model for satisfaction-related concepts. We used the model to categorize the concepts we found, and it can help the requirements engineer make an informed choice of variables when applying our method.

After compiling the list, we created a model of anticipated satisfaction and used it to derive the contents of a questionnaire for anticipated satisfaction. We refined the details of our approach in two empirical studies, which showed the feasibility of measuring anticipated satisfaction.

Our central result is MUSA, a method for measuring anticipated satisfaction. It is a questionnaire-based method in which a requirements engineer describes planned software features in a questionnaire and adds questions about the users' expected reaction to the features.

As a final contribution, we validated the method in two empirical studies and demonstrated its feasibility in a typical software development setting.

7.1 Conclusions

With these results, we make both theoretical and practical contributions to the field of software engineering. The list of satisfaction-related concepts together with the categorization method provides a comprehensive overview of the satisfaction measurement field, in which a multitude of different models are used.

The model of anticipated satisfaction allows for a more systematic understanding of how users react when asked to predict their reactions to a software, and what informs their answers. Since such predictions are inaccurate, it is important to analyze what influences them in order to build

methods which can correct for biases and other sources of low accuracy.

MUSA is a novel tool which allows measurement of *anticipated satisfaction* a concept which has not been described formally before in the scientific literature. It has practical applications in the software development process, especially as an instrument for early validation of requirements. It can be distinguished from other early validation approaches by its good scalability, few dependencies on other artefacts of the software development process (e.g. it does not require finished wireframes) and high flexibility. This makes it a viable alternative to existing approaches.

We consider MUSA to be directly applicable outside of an academic setting. Its requirements on resources, human skills, and organizational environment are low and likely to be met in the majority of commercial software development projects. It does not require excessive amounts of effort, and the timeline for organizing, conducting and evaluating the measurement can be easily incorporated in a typical development project, even when using rapid or agile methodologies. It has direct benefits for the software development project, as the detection of usability issues improves quality. By enabling this to happen early in the project, it also saves costs. These benefits, combined with the easily understood premise of the method and the familiarity with surveys among corporate executives, make the method easily defensible. In our experience, it is easy to gain support for applying the method and development teams appreciate the results it produces.

The validation showed that our method is generally useful, but has some shortcomings. Among those are the limited accuracy and thoroughness, and the requirement that the requirements engineer applying it has some knowledge of statistics to properly interpret and evaluate the results.

Despite these drawbacks, we consider our method to be a good choice for requirements validation. Its cost is moderate, it makes use of the best information source available, and it is highly scalable, unlike traditional validation methods such as interviews which are difficult to conduct with a large number of users.

The main focus of this dissertation is method development. While the method itself is intended for use in the software industry, our contributions also open new lines for theoretical and applied research.

The method itself can be further improved. While it delivers good quality results, our studies showed that the requirements engineers ask for more guidance especially in the selection of variables to use, and the participants find the anticipation step complicated. Research to address these issues can increase the method's adoption.

The variables used in the method correspond to the satisfaction related concepts we found in our literature study. In that study, we compiled evidence for the relation of 80 concepts with satisfaction, and did a metaanalysis on the 20 of them for which there was sufficient data available. To our knowledge, no other such extensive overviews of the field exist. The study showed

that, despite the multiple models proposed, it is not clearly understood how these variables relate to satisfaction, and that there is large variability in the empirical measurements even within the same concept. This has serious implications for the practice of satisfaction measurement, as it can result in unreliable measurements. More research is needed to understand the reasons behind the variability and develop more robust measurement methods.

Our work is the first to suggest the measurement of anticipated satisfaction, as opposed to actual satisfaction. We had to focus our work on a specific application, in this case the validation of requirements. Further work can explore its usefulness for other purposes, and build more precise models of anticipated satisfaction, to improve our understanding of it and find measurement approaches with higher validity.

The method presents a unique approach to requirements validation not only because of its use of anticipated satisfaction, but also because it is survey-based. It enriches the available set of requirements engineering tools, and more research is needed to investigate how it performs in relation to other requirements validation methods and under what circumstances it is the optimal choice. It can also serve as an example for researchers who wish to explore the potential of survey-based approaches in requirements engineering.

Part IV

Back matter

Bibliography

[1] ABELEIN, U. *User-Developer Communication in Large-Scale IT Projects*. Ph.d. thesis, Heidelberg university, Heidelberg, 2015. 62, 194

[2] ABELEIN, U., AND PAECH, B. Understanding the Influence of User Participation and Involvement on System Success–a Systematic Mapping Study. *Empirical Software Engineering* (2013). 4

[3] ACM. ACM digital library. http://dl.acm.org, 2016. Accessed: 19.08.2016. 44

[4] AL-BUSAIDI, K. A. An empirical investigation linking learners' adoption of blended learning to their intention of full e-learning. *Behaviour and Information Technology 32*, 11 (nov 2013), 1168–1176. 54, 158, 163, 177, 178, 182, 189

[5] AL-GAHTANI, S. S., AND KING, M. Attitudes, satisfaction and usage: Factors contributing to each in the acceptance of information technology. *Behaviour and Information Technology 18*, 4 (jan 1999), 277–297. 152, 158, 162, 163, 167, 187

[6] ALAPETITE, A., BOJE ANDERSEN, H., AND HERTZUM, M. Acceptance of speech recognition by physicians: A survey of expectations, experiences, and social influence. *International Journal of Human-Computer Studies 67*, 1 (jan 2009), 36–49. 150, 153, 154, 163, 165, 173, 178, 184, 186, 192

[7] ALEXANDER, I., AND BEUS-DUKIC, L. *Discovering Requirements*, 1 ed. Wiley, Chichester, 2009. 3

[8] ALGARNI, F., CHEUNG, Y., AND LEE, V. An empirical study of eMarketplaces customers' satisfaction: Evidence from Saudi Arabia. In *10th International Conference on Service Systems and Service Management* (jul 2013), IEEE, pp. 434–439. 180, 182, 192

[9] ANDERSON, R. E. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of marketing research* (1973), 38–44. 10

[10] AU, N., NGAI, E. W. T., AND CHENG, T. C. E. Extending the understanding of end user information systems satisfaction formation: an equitable needs fulfillment model approach. *MIS Quarterly 32*, 1 (mar 2008), 43–66. 168, 169, 187

[11] BAILEY, J. E., AND PEARSON, S. W. Development of a tool for measuring and analyzing computer user satisfaction. *Management science 29*, 5 (1983), 530–545. 11

[12] BARGAS-AVILA, J. A., LÖTSCHER, J., ORSINI, S., AND OPWIS, K. Intranet satisfaction questionnaire: Development and validationof a questionnaire to measure user satisfaction with the Intranet. *Computers in Human Behavior 25*, 6 (nov 2009), 1241–1250. 49, 54, 153, 160, 163, 165, 171, 179, 180, 185, 186, 191

[13] BARGAS-AVILA, J. A., ORSINI, S., DE VITO, M., AND OPWIS, K. ZeGo: Development and Validation of a Short Questionnaire to Measure User Satisfaction with e-Government Portals. *Advances in Human-Computer Interaction 2010* (jan 2010), 1–10. 49, 149, 155, 158, 159, 163, 166, 168, 179, 188, 191, 193

[14] BARKI, H. Determinants of user satisfaction judgements in information systems. In *Hawaii International Conference on System Sciences* (1990), pp. 408–417. 147, 148, 155, 160, 161, 163, 168, 170, 171, 179, 180, 185, 190, 193

[15] BAROUDI, J. J., OLSON, M. H., AND IVES, B. An empirical study of the impact of user involvement on system usage and information satisfaction. *Communications of the ACM 29*, 3 (mar 1986), 232–238. 192, 195

[16] BAROUDI, J. J., AND WANDA, J. O. A Short Form Measure of User Information Satisfaction: A Psychometric Evaluation and Notes on Use. *Journal of Management Information Systems 4*, 4 (1988), 44–59. 12

[17] BASILI, V. R., AND ROMBACH, D. H. The TAME project: Towards improvement-oriented software environments. *IEEE Transactions on software engineering 14*, 6 (1988), 758–773. 16

[18] BERGERON, F., RAYMOND, L., RIVARD, S., AND GARA, M.-F. Understanding EIS Use: An Empirical Test of a Behavioral Model. In *Hawaii International Conference on System Sciences* (1992). 170, 192

[19] BERNS, G. *Satisfaction: The science of finding true fulfillment*. Macmillan, Kindle version, 2010. 10

[20] BHATTACHERJEE, A. Understanding information systems continuance: an expectation-confirmation model. *MIS quarterly 25*, 3 (2001), 351–370. 10, 12, 56, 157, 158, 161, 193

[21] BIN, W., CHU-HONG, Z., QIONG-YU, H., ZHEN-PENG, L., HUANG, C.-H., LIAO, Q.-Y., AND COLLEGE, L. Empirical research on the factor of ERP's user customer satisfaction based on triadic reciprocal determinism. In *International Conference on Management Science and Engineering* (nov 2010), IEEE, pp. 58–66. 54, 158, 176, 182, 187, 192

[22] BOHLMANN, J. D., ROSA, J. A., BOLTON, R. N., AND QUALLS, W. J. The Effect of Group Interactions on Satisfaction Judgments: Satisfaction Escalation. *Marketing Science 25*, 4 (jul 2006), 301–321. 161, 165, 168, 184, 189, 195

[23] BRACE, I. *Questionnaire design: How to plan, structure and write survey material for effective market research*, 3 ed. Kogan Page, London, 2008. 84

[24] BROOKE, J. SUS - A quick and dirty usability scale. *Usability evaluation in industry 189*, 194 (1996), 1–8. 13

[25] BUYS, M., AND BROWN, I. Customer satisfaction with internet banking web sites: an empirical test and validation of a measuring instrument. In *Anual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries* (oct 2004), South African Institute for Computer Scientists and Information Technologists, pp. 44–52. 12, 156, 163, 182, 186

[26] CAPECE, G., AND CAMPISI, D. User satisfaction affecting the acceptance of an e-learning platform as a mean for the development of the human capital. *Behaviour and Information Technology 32*, 4 (apr 2013), 335–343. 10, 49, 163, 193

[27] CASALÓ, L., FLAVIÁN, C., AND GUINALÍU, M. The role of perceived usability, reputation, satisfaction and consumer familiarity on the website loyalty formation process. *Computers in Human Behavior 24*, 2 (mar 2008), 325–345. 10, 158, 175, 192

[28] CHANG, H. H., AND CHEN, S. W. The impact of customer interface quality, satisfaction and switching costs on e-loyalty: Internet experience as a moderator. *Computers in Human Behavior 24*, 6 (sep 2008), 2927–2944. 54, 148, 159, 160, 174, 175

[29] CHANG, L.-M., CHANG, S.-I., HO, C.-T., YEN, D. C., AND CHIANG, M.-C. Effects of IS characteristics on e-business success factors of small- and medium-sized enterprises. *Computers in Human Behavior 27*, 6 (nov 2011), 2129–2140. 168

[30] CHANG, P. K., AND CHONG, H. L. Customer satisfaction and loyalty on service provided by Malaysian telecommunication companies. In *International Conference on Electrical Engineering and Informatics* (jul 2011), IEEE, pp. 1–6. 158, 175, 183

[31] CHEN, C.-W. Impact of quality antecedents on taxpayer satisfaction with online tax-filing systems—An empirical study. *Information and Management 47*, 5-6 (aug 2010), 308–315. 173, 183, 187

[32] CHEN, C.-W. D., AND CHENG, C.-Y. J. Understanding consumer intention in online shopping: a respecification and validation of the DeLone and McLean model. *Behaviour and Information Technology 28*, 4 (jul 2009), 335–345. 158, 173, 183, 187, 192, 195

[33] CHEN, S. C., AND CHEN, H. H. The empirical study of customer satisfaction and con-

tinued behavioural intention towards self-service banking: technology readiness as an antecedent. *International Journal of Electronic Finance 3*, 1 (mar 2009), 64. 154, 158, 182

[34] CHEN, S.-C., YEN, D. C., AND HWANG, M. I. Factors influencing the continuance intention to the usage of Web 2.0: An empirical study. *Computers in Human Behavior 28*, 3 (may 2012), 933–941. 158, 179, 185

[35] CHEN, Y.-Y., HUANG, H.-L., HUANG, W.-N., AND SUNG, S.-F. Confirmation of Expectations and Satisfaction with an On-Line Service: The Role of Internet Self-Efficacy. In *International Conference on New Trends in Information and Service Science* (jun 2009), IEEE, pp. 880–885. 158, 161, 193

[36] CHEUNG, C. M. K., AND LEE, M. K. O. What drives members to continue sharing knowledge in a virtual professional community? the role of knowledge self-efficacy and satisfaction. In *International Conference on Knowledge Science, Engineering and Management* (nov 2007), pp. 472–484. 158, 161

[37] CHEUNG, C. M. K., AND LEE, M. K. O. The structure of Web-based information systems satisfaction: Testing of competing models. *Journal of the American Society for Information Science and Technology 59*, 10 (aug 2008), 1617–1630. 147, 191

[38] CHEUNG, C. M. K., AND LEE, M. K. O. User satisfaction with an internet-based portal: An asymmetric and nonlinear approach. *Journal of the American Society for Information Science and Technology 60*, 1 (jan 2009), 111–122. 148, 155, 160

[39] CHIN, J. P., DIEHL, V. A., NORMAN, K. L., AND NORMAN, L. K. Development of an instrument measuring user satisfaction of the human-computer interface. In *SIGCHI conference on Human factors in computing systems* (New York, New York, USA, may 1988), ACM Press, pp. 213–218. 13

[40] CHOI, J. H., AND LEE, H.-J. Facets of simplicity for the smartphone interface: A structural model. *International Journal of Human-Computer Studies 70*, 2 (feb 2012), 129–142. 12, 149, 156

[41] CHOU, S.-W., AND CHEN, P.-Y. The influence of individual differences on continuance intentions of enterprise resource planning (ERP). *International Journal of Human-Computer Studies 67*, 6 (jun 2009), 484–496. 150, 158, 182

[42] CHUNG, N., AND KWON, S. J. Effect of trust level on mobile banking satisfaction: a multi-group analysis of information system success instruments. *Behaviour & Information Technology 28*, 6 (nov 2009), 549–562. 171, 173, 187

[43] COHN, M. *User stories applied*, 1 ed. Addison-Wesley Professional, Boston, 2004. 22

[44] COOHAROJANANONE, N., CHOFA, S., AND PHIMOLTARES, S. A Study on Intention to Use Factor in the Internet Banking Websites in Thailand. In *IEEE/IPSJ International Symposium on Applications and the Internet* (jul 2011), IEEE, pp. 556–561. 158

[45] COOPER, H. *Research Synthesis and Meta-Analysis*, 5 ed. Sage, Kindle version, 2016. 15, 52, 58, 62, 63, 64

[46] COURSARIS, C. K., HASSANEIN, K., HEAD, M. M., AND BONTIS, N. The impact of distractions on the usability and intention to use mobile devices for wireless data services. *Computers in Human Behavior 28*, 4 (jul 2012), 1439–1449. 151, 158, 161, 164, 165

[47] COX, A., AND FISHER, M. An Expectation-Based Model of Web Search Behaviour. In *Second International Conferences on Advances in Computer-Human Interactions* (feb 2009), pp. 49–56. 161, 168, 189

[48] DAI, C. C.-Y., KAO, M. M.-T., HARN, C.-T. C., YUAN, Y.-H., AND CHEN, W.-F. The research on user satisfaction of easy teaching Web of Taipei assessed via information quality, system quality, and Technology Acceptance Model. In *Computer Science and education* (aug 2011), IEEE, pp. 758–762. 152, 163, 173, 187, 193

[49] DAMASIO, A. *Self comes to mind*, 1 ed. Pantheon, New York, 2010. 11

[50] DAVIS, F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly 13*, 3 (1989), 319–340. 12

[51] DELONE, W., AND MCLEAN, E. Information systems success revisited. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (2002), pp. 1–11. 3, 12, 61

[52] DELONE, W. H., AND MCLEAN, E. R. Information systems success: The quest for the dependent variable. *Information Systems Research 3*, 1 (1992), 60–95. 12

[53] DIENER, E. D., EMMONS, R. A., LARSEN, R. J., AND GRIFFIN, S. The satisfaction with life scale. *Journal of personality assessment 49*, 1 (1985), 71–75. 9

[54] DIX, A., FINLEY, J., ABOWD, G., AND BEALE, R. *Human-computer interaction*, 3 ed. Prentice Hall, Harlow, 2004. 38

[55] DOLL, W. J., AND TORKZADEH, G. The Measurement of End-User Computing Satisfaction. *MIS Quarterly 12*, 2 (jun 1988), 259. 12, 82, 159

[56] DOLL, W. J., XIA, W., AND TORKZADEH, G. A Confirmatory Factor Analysis of the End-User Computing Satisfaction Instrument. *MIS Quarterly 18*, 4 (dec 1994), 453. 54, 61

[57] DZIKOVSKA, M. O., MOORE, J. D., STEINHAUSER, N., AND CAMPBELL, G. Exploring User

Satisfaction in a Tutorial Dialogue System. In *SIGDIAL Conference* (jun 2011), Association for Computational Linguistics, pp. 162–172. 54, 189

[58] EHLERS, E. G. States of matter. http://www.britannica.com/science/phase-state-of-matter. In: Encyclopedia Britannica online, accessed: 13.07.2016. 13

[59] EKMAN, P. *Emotions Revealed: Understanding Faces and Feelings*, 1 ed. Holt, Kindle version, 2012. 11

[60] FANG-MEI TSENG, HSIN-YEN CHIANG, HUIYI LO, TSENG, F.-M., CHIANG, H.-Y., AND LO, H. Stating mobile phone upgrading behavior. In *Technology Management in the Energy Smart World (PICMET)* (2011), pp. 1–10. 12

[61] FENG, J., AND SEARS, A. Beyond errors: measuring reliability for error-prone interaction devices. *Behaviour and Information Technology 29*, 2 (mar 2010), 149–163. 168, 180

[62] FENTON, N. Software measurement: a necessary scientific basis. *IEEE Transactions of Software Engineering 20*, 3 (1994), 199–206. 10

[63] FENTON, N., AND BIEMAN, J. *Software metrics: A Rigorous and Practical Approach*, 3 ed. CRC Press, Boca Raton, 2014. 16

[64] FINSTAD, K. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies 5*, 3 (2010), 104–110. 80

[65] FINSTAD, K. The Usability Metric for User Experience. *Interacting with Computers 22*, 5 (sep 2010), 323–327. 192

[66] FOERSTER, H. V. Abbau und Aufbau. In *Lebende Systeme*, F. Simon, Ed. Springer, Berlin, 1988, pp. 19–33. 13

[67] FRIGG, R., AND HARTMANN, S. Models in Science. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Stanford university, 2012. Accessed: 22.10.2016. 13

[68] FROKJAER, E., HERTZUM, M., AND HORNBÆK, K. Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? In *SIGCHI conference on Human factors in computing systems* (New York, New York, USA, apr 2000), ACM Press, pp. 345–352. 10, 164

[69] GARCIA, F., BERTOA, M. F., CALERO, C., VALLECILLO, A., RUIZ, F., PIATTINI, M., AND GENERO, F. Towards a consistent terminology for software measurement. *Information and Software Technology 48* (2006), 631–644. 9, 16

[70] GEDIGA, G., HAMBORG, K. C. K.-C., AND DÜNTSCH, I. The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of

software systems. *Behaviour and Information Technology 18*, 3 (1999), 151–164. 13

[71] GIERE, R. N. How Models Are Used to Represent Reality. *Philosophy of Science 71*, 5 (2004), 742–752. 13

[72] GOLDSTEIN, D. K., AND ROCKART, J. F. An Examination of Work-Related Correlates of Job Satisfaction in Programmer/Analysts. *MIS Quarterly 8*, 2 (jun 1984), 103. 9

[73] GRAY, D. *Doing research in the real world*, 1 ed. Sage, London, 2004. 67, 68, 80, 85, 86

[74] GREEN, D. T., AND PEARSON, J. M. Integrating website usability with the electronic commerce acceptance model. *Behaviour and Information Technology 30*, 2 (mar 2011), 181–199. 158, 159, 163, 181, 193

[75] GUADAGNO, R. E., SWINTH, K. R., AND BLASCOVICH, J. Social evaluations of embodied agents and avatars. *Computers in Human Behavior 27*, 6 (nov 2011), 2380–2385. 152, 154, 166, 167, 186, 190

[76] GUDIGANTALA, N., SONG, J., AND JONES, D. User satisfaction with Web-based DSS: The role of cognitive antecedents. *International Journal of Information Management 31*, 4 (aug 2011), 327–338. 148, 164, 165

[77] HALILOVIC, S., AND CICIC, M. Understanding determinants of information systems users' behaviour: a comparison of two models in the context of integrated accounting and budgeting software. *Behaviour and Information Technology 32*, 12 (dec 2013), 1280–1291. 158, 161, 193

[78] HAND, D. J. *Measurement Theory and Practice: The World Through Quantification*, 1 ed. Wiley-Blackwell, Chichester, 2010. 10

[79] HARRISON, A. W., AND JR, K. R. A General Measure of User Computing Satisfaction. *Computers in Human Behavior 12*, 1 (1996), 79–92. 150, 166, 189, 191, 192

[80] HARTSON, R., ANDRE, T. S., AND WILLIGES, R. C. Criteria For Evaluating Usability Evaluation Methods. *International journal of human-computer interaction 13*, 4 (dec 2001), 373–410. 95, 110

[81] HARTWICK, J. Delineating the dimensions of user participation: a replication and extension. In *Hawaii International Conference on System Sciences* (1997). 195

[82] HASSANEIN, K., HEAD, M., AND WANG, F. Understanding Student Satisfaction in a Mobile Learning Environment: The Role of Internal and External Facilitators. In *International Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable (ICMB-GMR)* (jun 2010), IEEE, pp. 289–296. 167, 193

[83] HASSENZAHL, M., SCHÖBEL, M., AND TRAUTMANN, T. How motivational orientation influences the evaluation and choice of hedonic and pragmatic interactive products: The role of regulatory focus. *Interacting with Computers 20*, 4-5 (sep 2008), 473–479. 148

[84] HAU, Y. S., KIM, G., AND KIM, B. Antecedents of user satisfaction in the context of mobile data services: the moderating role of variety and rate of usage. *International Journal of Mobile Communications 10*, 6 (oct 2012), 617. 147, 148, 155, 160, 167, 171, 173, 177, 180, 187, 190, 193

[85] HAWK, S. R., AND DOS SANTOS, B. L. Successful System Development: The Effect of Situational Factors on Alternate User Roles. *IEEE Transactions on Engineering Management 38*, 4 (1991), 316–327. 195

[86] HERNANDEZ, B., JIMENEZ, J., AND JOSE MARTIN, M. The impact of self-efficacy, ease of use and usefulness on e-purchasing: An analysis of experienced e-shoppers. *Interacting with Computers 21*, 1-2 (jan 2009), 146–156. 152, 158, 163, 182, 192, 193

[87] HIGGINS, J. P. T., THOMPSON, S. G., AND SPIEGELHALTER, D. J. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A: Statistics in Society 172*, 1 (2009), 137–159. 58, 64

[88] HINTZE, J. L., AND NELSON, R. D. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician 52*, 2 (may 1998), 181–184. 93

[89] HOFMANN, H. F., AND LEHNER, F. Requirements engineering as a success factor in software projects. *IEEE Software 18*, 4 (2001), 58–66. 3

[90] HOLM, S. A simple sequentially rejective multiple test procedure. *Statistics 6*, 2 (1979), 65–70. 59

[91] HONG, S., AND KIM, J. Architectural criteria for website evaluation – conceptual framework and empirical validation. *Behaviour and Information Technology 23*, 5 (sep 2004), 337–357. 49, 149

[92] HSU, M.-H., YEN, C.-H., CHIU, C.-M., AND CHANG, C.-M. A longitudinal investigation of continued online shopping behavior: An extension of the theory of planned behavior. *International Journal of Human-Computer Studies 64*, 9 (sep 2006), 889–904. 152, 153, 158, 161, 184

[93] HULLEY, S. B., CUMMINGS, S. R., BROWNER, W. S., GRADY, D. G., AND NEWMAN, T. B. *Designing clinical research*, 4 ed. Lippincott Williams and Wilkins, Philadelphia, 2013. 49

[94] IEEE. IEEE Xplore. http://ieeexplore.ieee.org/, 2016. Accessed: 12.09.2016. 44

[95] IGBARIA, M. The impact of user attitudes toward microcomputer usage on system usage

and user satisfaction. *ACM SIGCPR Computer Personnel 12*, 2 (dec 1989), 15–21. 152, 158

[96] IGBARIA, M., SCHIFFMAN, S. J., AND WIECKOWSKI, T. J. The respective roles of perceived usefulness and perceived fun in the acceptance of microcomputer technology. *Behaviour and Information Technology Information Technology 13*, 6 (nov 1994), 349–361. 150, 167, 188, 192, 193

[97] IN, K. J., HYE, A. S., AND LEE, C. C. A study on service quality determinants that influence continued success of ortal online community services. In *IEEE International Conference on Computer Science and Information Technology* (2009), IEEE, pp. 171–175. 151, 166, 179, 180, 181

[98] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability, 1998. 12

[99] ITU-T. ITU-T Recommendation G.109 Amendment 1, 2007. 11

[100] IVES, B., OLSON, M. H., AND BAROUDI, J. J. User Satisfaction Measurement Methodologies: Extending the User Satisfaction Questionnaire. *Communications of the ACM 26*, 10 (aug 1983), 1008–1012. 10, 12

[101] JIANG, X. Enhancing Users' Continuance Intention to E-Government Portals: An Empirical Study. In *2011 International Conference on Management and Service Science* (aug 2011), IEEE, pp. 1–4. 158, 173, 180, 181, 182

[102] JIN, B. S., YOON, S. H., AND JI, Y. G. Development of a Continuous Usage Model for the Adoption and Continuous Usage of a Smartphone. *International Journal of Human-Computer Interaction 29*, 9 (sep 2013), 563–581. 154, 158, 161, 163, 167, 168, 178, 183, 184, 192, 193, 195

[103] JIN, X.-L., LEE, M. K. O., AND CHEUNG, C. M. K. Predicting continuance in online communities: model development and empirical test. *Behaviour and Information Technology 29*, 4 (jul 2010), 383–394. 154, 158, 161

[104] JOSHI, K., PERKINS, W. C., AND BOSTROM, R. P. Some new factors influencing user information satisfaction. In *Computer personnel research conference - CPR '86* (New York, New York, USA, oct 1986), ACM Press, pp. 27–42. 172, 173, 186, 194

[105] KANG, C.-R., HUNG, M.-C., YANG, S.-T., HSIEH, T.-C., AND TANG, S.-M. Factors affecting the continued intention of mobile shopping. In *IEEE International Conference on Industrial Engineering and Engineering Management* (2010), Ieee, pp. 710–713. 150, 158, 167, 172, 173, 187, 193

[106] KARJALUOTO, H., JARVENPAA, L., AND KAUPPI, V. Antecedents of online banking satisfaction and loyalty: empirical evidence from Finland. *International Journal of Electronic Finance 3*, 3 (aug 2009), 253. 163, 175, 180, 182, 193

[107] KASSIM, E. S., AND HUSSIN, H. User attitude, organizational learning and dynamic capability in government-to-business success. In *IEEE International Conference on Management of Innovation and Technology* (2010), IEEE, pp. 1061–1066. 152, 162, 165, 173, 192

[108] KIM, J., HONG, S., MIN, J., AND LEE, H. Antecedents of application service continuance: A synthesis of satisfaction and trust. *Expert Systems with Applications 38*, 8 (aug 2011), 9530–9542. 12, 158, 173, 183, 187, 190

[109] KIM, Y. H., KIM, D. J., AND WACHTER, K. A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision Support Systems 56* (dec 2013), 361–370. 158, 195

[110] KIRAKOWSKI, J., AND CLARIDGE, N. WAMMI. http://www.wammi.com, 2016. Accessed: 26.08.2016. 13

[111] KIRAKOWSKI, J., AND CORBETT, M. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology 24*, 3 (sep 1993), 210–212. 13

[112] KITCHENHAM, B. A., AND CHARTERS, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering (Version 2.3). Tech. Rep. EBSE 2007-001, Keele University; University of Durham, Keele, Staffs, UK; Durham, UK, 2007. 41, 48, 64

[113] KLUGE, A., GRAUEL, B., AND BURKOLTER, D. Combining principles of Cognitive Load Theory and diagnostic error analysis for designing job aids: Effects on motivation and diagnostic performance in a process control task. *Applied ergonomics 44*, 2 (mar 2013), 285–296. 173

[114] KNUUTTILA, T. *Models as epistemic artefacts: Toward a non-representationalist account of scientific representation*. PhD thesis, University of Helsinki, 2005. 13

[115] KOIVUMÄKI, T., RISTOLA, A., AND KESTI, M. The effects of information quality of mobile information services on user satisfaction and service acceptance–empirical evidence from Finland. *Behaviour & Information Technology 27*, 5 (sep 2008), 375–385. 156, 158

[116] KONRADT, U., CHRISTOPHERSEN, T., AND SCHAEFFER-KUELZ, U. Predicting user satisfaction, strain and system usage of employee self-services. *International Journal of Human-Computer Studies 64*, 11 (nov 2006), 1141–1153. 51, 149, 163, 164, 165, 186, 189, 192, 193, 194

[117] KOO, C., AND WATI, Y. E-Healthcare Service: An Investigation of the Antecedents, Moderating Roles, and Consequences. In *Hawaii International Conference on System Sciences*

(jan 2011), IEEE, pp. 1–10. 158, 159, 187

[118] KOTONYA, G., AND SOMMERVILLE, I. *Requirements engineering: processes and techniques*, 1 ed. Wiley Publishing, Hoboken, 1998. 3

[119] KUO, Y.-F., WU, C.-M., AND DENG, W.-J. The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Computers in Human Behavior 25*, 4 (jul 2009), 887–896. 183, 195

[120] LAI, C.-Y., AND YANG, H.-L. The reasons why people continue editing wikipedia content — task value confirmation perspective. *Behaviour and Information Technology 33*, 12 (may 2014), 1371–1382. 154, 158, 161, 169, 174, 195

[121] LAI, J.-Y. Assessment of employees' perceptions of service quality and satisfaction with e-business. *International Journal of Human-Computer Studies 64*, 9 (apr 2006), 926–938. 150, 151, 166, 180, 181

[122] LAING, G. J. The Classification of Models A Proposal. *Interdisciplinary Science Reviews 6*, 4 (1981), 355–363. 13

[123] LAUESEN, S. *User Interface Design - A software engineering perspective*, 1 ed. Pearson education, Harlow, 2005. 22, 83

[124] LAUESEN, S. Why the electronic land registry failed. In *REFSQ'12* (2012), Springer, pp. 1–15. 3

[125] LAW, E., ROTO, V., HASSENZAHL, M., VERMEEREN, A., AND KORT, J. Understanding, Scoping and Defining User eXperience: A Survey Approach. In *Conference on Human Factors in Computing Systems* (2009), pp. 719–728. 12

[126] LEE, K. C., AND CHUNG, N. Understanding factors affecting trust in and satisfaction with mobile banking in Korea: A modified DeLone and McLean's model perspective. *Interacting with Computers 21*, 5-6 (dec 2009), 385–392. 160, 173, 187, 190

[127] LEE, S. Understanding User Experience with Computer-Based Applications with Different Use Purposes. *International Journal of Human-Computer Interaction 29*, 11 (nov 2013), 689–701. 56, 149, 192

[128] LEITE, R. S., DE CARVALHO, R. B., AND FILHO, C. G. Measuring Perceived Quality and Satisfaction of ERP Systems: An Empirical Study with Customers of a Brazilian Software Company. In *Hawaii International Conference on System Sciences* (2009), IEEE, pp. 1–8. 158, 168, 175, 195

[129] LEWIS, J. R. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction 14*, 3-4 (sep 2002),

463–488. 13

[130] LIAO, C., CHEN, J.-L., AND YEN, D. C. Theory of planning behavior (TPB) and customer satisfaction in the continued use of e-service: An integrated model. *Computers in Human Behavior 23*, 6 (nov 2007), 2804–2822. 12, 152, 153, 158, 161, 163, 185, 193

[131] LIAW, S.-S., CHANG, W.-C., HUNG, W.-H., AND HUANG, H.-M. Attitudes toward search engines as a learning assisted tool: approach of Liaw and Huang's research model. *Computers in Human Behavior 22*, 2 (mar 2006), 177–190. 158, 167, 170, 182, 189

[132] LIM, J.-S., AL-AALI, A., HEINRICHS, J. H., AND LIM, K.-S. Testing alternative models of individuals' social media involvement and satisfaction. *Computers in Human Behavior 29*, 6 (nov 2013), 2816–2828. 158, 193, 195

[133] LIMAYEM, M., AND CHEUNG, C. M. K. Predicting the continued use of Internet-based learning technologies: the role of habit. *Behaviour and Information Technology 30*, 1 (jan 2011), 91–99. 158, 161, 172, 192, 193

[134] LIN, H., FAN, W., WALLACE, L., AND ZHANG, Z. An Empirical Study of Web-Based Knowledge Community Success. In *Hawaii International Conference on System Sciences (HICSS'07)* (2007), pp. 178c—-178c. 173, 185, 187, 192

[135] LIN, H.-F., AND LEE, G.-G. Determinants of success for online communities: an empirical study. *Behaviour & Information Technology 25*, 6 (nov 2006), 479–488. 158, 173, 175, 183, 186, 187

[136] LIN, J.-S. C., AND HSIEH, P.-L. The influence of technology readiness on satisfaction and behavioral intentions toward self-service technologies. *Computers in Human Behavior 23*, 3 (may 2007), 1597–1615. 158, 178

[137] LIN, K.-M., CHEN, N.-S., AND FANG, K. Understanding e-learning continuance intention: a negative critical incidents perspective. *Behaviour and Information Technology 30*, 1 (jan 2011), 77–89. 152, 158, 163, 168, 193

[138] LIN, W.-S. Perceived fit and satisfaction on web learning performance: IS continuance intention and task-technology fit perspectives. *International Journal of Human-Computer Studies 70*, 7 (jul 2012), 498–507. 49, 50, 51, 52, 53, 56, 158, 189

[139] LIU, Y., CHEN, Y., AND ZHOU, C. Determinants Affecting End-User Satisfaction of Information Technology Service. In *International Conference on Service Systems and Service Management* (oct 2006), Ieee, pp. 478–481. 153, 163, 168, 183, 193

[140] LOCKE, E. A. What is job satisfaction? *Organizational Behavior and Human Performance 4*, 4 (nov 1969), 309–336. 9

[141] LOKE, S.-P., NOOR, N. M., AND KHALID, K. Customer satisfaction towards internet banking services: Case analysis on a Malaysian bank. In *IEEE Colloquium on Humanities, Science and Engineering (CHUSER)* (dec 2012), pp. 159–163. 51, 149, 164, 176, 182, 186

[142] LOWRY, P. B., SPAULDING, T., WELLS, T., MOODY, G., MOFFIT, K., AND MADARIAGA, S. A Theoretical Model and Empirical Results Linking Website Interactivity and Usability Satisfaction. In *Hawaii International Conference on System Sciences (HICSS'06)* (jan 2006), vol. 6, IEEE, pp. 123a—-123a. 161, 168, 174

[143] MAHMOOD, M. A. M., BURN, J. M., GEMOETS, L. A., JACQUEZ, C., AND JAQUEZ, C. Variables affecting information technology end-user satisfaction: a meta-analysis of the empirical literature. *International Journal of Human-Computer Studies 52*, 4 (apr 2000), 751–771. 64

[144] MÄNTYMÄKI, M., AND ISLAM, A. N. Social virtual world continuance among teens: uncovering the moderating role of perceived aggregate network exposure. *Behaviour and Information Technology 33*, 5 (jan 2014), 536–547. 49, 158, 161, 167, 193

[145] MARTILLA, J., AND JAMES, J. Importance-performance analysis. *The journal of marketing 41*, 1 (1977), 77–79. 22

[146] MCHANEY, R., HIGHTOWER, R., AND PEARSON, J. A validation of the end-user computing satisfaction instrument in Taiwan. *Information and Management 39*, 6 (may 2002), 503–511. 148, 156, 163, 171, 190

[147] Merriam-Webster.com. http://www.merriam-webster.com/dictionary/, 2013. Accessed: 14.08.2013. 145, 147, 149, 153, 154, 155, 165, 169, 170, 171, 173, 179, 185, 188

[148] MONNICKENDAM, M., SAVAYA, R., AND WAYSMAN, M. Targeting implementation efforts for maximum satisfaction with new computer systems: Results from four human service agencies. *Computers in Human Behavior 24*, 4 (jul 2008), 1724–1740. 167, 176, 178, 185, 186, 193, 194, 195

[149] NORDTVEDT, K. L. Gravity. http://www.britannica.com/science/gravity-physics, 2016. Encyclopedia Britannica Online. Accessed: 22.06.2016. 14

[150] NWANKPA, J., AND ROUMANI, Y. Understanding the link between organizational learning capability and ERP system usage: An empirical examination. *Computers in Human Behavior 33* (apr 2014), 224–234. 162, 176, 192

[151] OGARA, S. O., KOH, C. E., AND PRYBUTOK, V. R. Investigating factors affecting social presence and user satisfaction with Mobile Instant Messaging. *Computers in Human Behavior 36* (jul 2014), 453–459. 55, 170, 177, 184

[152] OLIVEIRA, R. D., CHERUBINI, M., AND OLIVER, N. Influence of personality on satisfaction with mobile phone services. *ACM Transactions on Computer-Human Interaction 20*, 2 (may 2013), 1–23. 173, 192

[153] OLIVER, R. *Satisfaction: A behavioral perspective on the consumer*, 2 ed. M. E. Sharpe, Armonk, 2010. 10, 12, 26, 55, 89, 90, 110, 169

[154] OZEN, C., AND BASOGLU, N. Impact of Man-Machine Interaction Factors on Enterprise Resource Planning (ERP) Software Design. In *Technology Management for the Global Future - PICMET 2006 Conference* (jul 2006), Ieee, pp. 2335–2341. 163, 193

[155] PENG, H., SONG, H., ZHANG, Z., CHEN, Y., ZOU, X., AND XIAO, L. A Study on User Experience of Online Games. In *WRI World Congress on Software Engineering* (2009), pp. 185–189. 186

[156] POHL, K. *Requirements engineering: fundamentals, principles, and techniques*. Springer, Berlin, 2010. 3

[157] PROYNOVA, R., AND PAECH, B. Do Stakeholders Understand Feature Descriptions? A Live Experiment. In *REFSQ* (2012), pp. 265–280. 20

[158] PROYNOVA, R., AND PAECH, B. Factors influencing user feedback on predicted satisfaction with software systems. In *19th International Working Conference on Requirements Engineering: Foundation for Software Quality* (2013). 30

[159] REDZUAN, F., HASSIM, N., MALAYSIAN, R., AND FORCE, A. Usability study on Integrated Computer Management System for Royal Malaysian Air Force (RMAF). In *IEEE Conference on e-Learning, e-Management and e-Services* (dec 2013), IEEE, pp. 93–99. 165, 168, 174

[160] RIVARD, S., AND HUFF, S. L. Factors of success for end-user computing. *Communications of the ACM 31*, 5 (may 1988), 552–561. 152

[161] RUPP, C. *Requirements Engineering und Management*, 5 ed. Hanser, Vienna, 2009. 3, 31

[162] SANCHEZ-FRANCO, M., AND RONDAN-CATALUÑA, F. J. Connection between customer emotions and relationship quality in online music services. *Behaviour and Information Technology 29*, 6 (nov 2010), 633–651. 154, 166, 190

[163] SAQI, S. B. *Requirements Validation Techniques practiced in industry : Studies of six companies*. PhD thesis, Blekinge institute of technology, 2008. 3

[164] SARIS, W. E., AND GALLHOFER, I. N. *Design, evaluation, and analysis of questionnaires for survey research*, 1 ed., vol. 548. Wiley, Hoboken, 2007. 67, 80, 86

[165] SAURO, J. 17 Periodicals For Usability Research. http://www.measuringusability.com/

blog/usability-periodicals.php, 2013. Accessed: 11.04.2013. 44

[166] SCHALL, A. The future of UX research: uncovering the true emotions of our users. *User Experience Magazine 15*, 2 (2015). 11

[167] SHI, N., LEE, M. K. O., CHEUNG, C. M. K., AND CHEN, H. The continuance of online social networks: how to keep people using Facebook? In *Hawaii International Conference on System Sciences* (2010), IEEE, pp. 1–10. 158, 161

[168] SHIAU, W.-L., AND LUO, M. M. Continuance intention of blog users: the impact of perceived enjoyment, habit, user involvement and blogging time. *Behaviour and Information Technology 32*, 6 (jun 2013), 570–583. 158, 161, 167, 172, 195

[169] SHIPPS, B., AND PHILLIPS, B. Social Networks, Interactivity and Satisfaction: Assessing Socio-Technical Behavioral Factors as an Extension to Technology Acceptance. *Journal of theoretical and applied electronic commerce research 8*, 1 (apr 2013), 7–8. 152, 153, 163, 171, 176, 177, 193, 194

[170] SOMMERVILLE, I. *Software Engineering*, 8 ed. Pearson, Harlow, 2004. 39

[171] SØREBØ, Ø., AND EIKEBROKK, T. R. Explaining IS continuance in environments where usage is mandatory. *Computers in Human Behavior 24*, 5 (sep 2008), 2357–2371. 161, 163, 193

[172] STONE, R. W., AND BAKER-EVELETH, L. Students' expectation, confirmation, and continuance intention to use electronic textbooks. *Computers in Human Behavior 29*, 3 (may 2013), 984–990. 158, 161, 193

[173] SUSANTO, A., BAHAWERES, R. B., AND ZO, H. Exploring the influential antecedents of actual use of internet banking services in Indonesia. In *IEEE Conference on Control, Systems and Industrial Informatics* (sep 2012), IEEE, pp. 244–249. 182, 183, 187, 190, 192

[174] TENG, C.-I. Customization, immersion satisfaction, and online gamer loyalty. *Computers in Human Behavior 26*, 6 (nov 2010), 1547–1554. 145, 148, 175

[175] THE EDITORS OF ENCYCLOPEDIA BRITANNICA. Scientific theory. http://www.britannica.com/science/scientific-theory. Encyclopedia Britannica online, Accessed: 17.03.2016. 13

[176] THONG, J. Y. L., HONG, S.-J., AND TAM, K. Y. The effects of post-adoption beliefs on the expectation-confirmation model for information technology continuance. *International Journal of Human-Computer Studies 64*, 9 (sep 2006), 799–810. 158, 161, 163, 167, 193

[177] TOURANGEAU, R., RIPS, L. J., AND RASINSKI, K. *The psychology of survey response*, 1 ed. Cambridge University Press, Cambridge, 2000. 19, 80, 90

[178] TUFTE, E. R. *Envisioning information.*, 1 ed. Graphics press, Cheshire, 1990. 90

[179] TULLIS, T., AND ALBERT, W. *Measuring the user experience*, 2 ed. Morgan Kaufmann, Kindle version, 2013. 11, 16, 89, 145, 174, 191

[180] UDO, G. J., BAGCHI, K. K., AND KIRS, P. J. Using SERVQUAL to assess the quality of e-learning experience. *Computers in Human Behavior 27*, 3 (may 2011), 1272–1283. 158, 180, 189

[181] UDO, G. J., BAGCHI, K. K., AND KIRS, P. J. Exploring the role of espoused values on e-service adoption: A comparative analysis of the US and Nigerian users. *Computers in Human Behavior 28*, 5 (sep 2012), 1768–1781. 10, 158, 163, 187, 193

[182] VACHIRAPORN, K., RACTHAM, P., AND KHAYUN, V. Measuring e-excise tax success factors: applying the DeLone and McLean information systems success model. In *Hawaii International Conference on.* (jan 2011), IEEE, pp. 1–10. 153, 173, 183, 187, 192

[183] VENKATESH, V., AND BALA, H. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences 39*, 2 (2008), 273. 12, 14, 61

[184] VENKATESH, V., MORRIS, M., DAVIS, G., AND DAVIS, F. User acceptance of information technology: Toward a unified view. *MIS quarterly 27*, 3 (2003), 425–478. 12

[185] VERHAGEN, T., FELDBERG, F., VAN DEN HOOFF, B., MEENTS, S., AND MERIKIVI, J. Satisfaction with virtual worlds: An integrated model of experiential value. *Information and Management 48*, 6 (aug 2011), 201–207. 10, 153, 163

[186] VIECHTBAUER, W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software 58*, 9 (2014). 58

[187] WANG, C. Antecedents and consequences of perceived value in Mobile Government continuance use: An empirical research in China. *Computers in Human Behavior 34* (may 2014), 140–147. 181, 195

[188] WANG, W., HSIEH, J. P.-A., AND SONG, B. Understanding User Satisfaction With Instant Messaging: An Empirical Survey Study. *International Journal of Human-Computer Interaction 28*, 3 (mar 2012), 153–162. 167, 177, 184, 193

[189] WANG, W.-T., AND LAI, Y.-J. Examining the adoption of KMS in organizations from an integrated perspective of technology, individual, and organization. *Computers in Human Behavior 38* (sep 2014), 55–67. 158, 183, 187

[190] WARE, J. E., SNYDER, M. K., WRIGHT, W. R., AND DAVIES, A. R. Defining and measuring patient satisfaction with medical care. *Evaluation and program planning 6*, 3 (1983), 247–263. 9

[191] WEI-CHEN, T. The impact of customer' s affective trait on e-service quality and satisfaction - travel website cases. In *ICACT* (2011), pp. 1434–1439. 155, 166, 173, 179, 187

[192] WEISBERG, M. *Simulation and similarity: Using models to understand the world*, 1 ed. Oxford University Press, Kindle version, 2012. 13

[193] WIEBE, E. N., LAMB, A., HARDY, M., AND SHAREK, D. Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior 32* (mar 2014), 123–132. 149, 151, 171, 192

[194] WILLIAMSON, D. F., PARKER, R. A., AND KENDRICK, J. S. The box plot: A simple visual method to interpret data. *Annals of Internal Medicine 110*, 11 (1989), 916–921. 93

[195] WIXOM, B. H., AND TODD, P. A. A Theoretical Integration of User Satisfaction and Technology Acceptance. *Information Systems Research 16*, 1 (mar 2005), 85–102. 12, 147, 148, 152, 155, 158, 160, 163, 170, 171, 173, 180, 187, 190, 193

[196] WU, I.-L., AND HUANG, C.-Y. Analysing complaint intentions in online shopping: the antecedents of justice and technology use and the mediator of customer satisfaction. *Behaviour and Information Technology* (jan 2014), 1–12. 155, 161, 169, 193

[197] WU, Y.-L., TAO, Y.-H., LI, C.-P., WANG, S.-Y., AND CHIU, C.-Y. User-switching behavior in social network sites: A model perspective with drill-down analyses. *Computers in Human Behavior 33* (apr 2014), 92–103. 153, 158, 165, 181, 187, 189

[198] XIA, W., AND LEE, G. Grasping the complexity of IS development projects. *Communications of the ACM 47*, 5 (may 2004), 68–74. 156

[199] YANG, X.-C., ZHANG, X.-H., AND ZUO, F. Word of mouth: the effects of marketing efforts and customer satisfaction. In *International Joint Conference on Artificial Intelligence* (apr 2009), pp. 687–690. 176, 184

[200] YEO, J. S. J., AURUM, A., HANDZIC, M., AND PARKIN, P. When Technology is Mandatory—Factors Influencing Users Satisfaction. In *Conference on Computers in Education* (dec 2002), p. 1023. 163, 193

[201] YOON, C. Antecedents of customer satisfaction with online banking in China: The effects of experience. *Computers in Human Behavior 26*, 6 (nov 2010), 1296–1304. 156, 160, 163, 182, 185, 186

[202] YUNG-MING LI, AND YUNG-SHAO YEN. Service quality's impact on mobile satisfaction and intention to use 3G service. In *Hawaii International Conference on System Sciences* (2009), pp. 1–10. 148, 158, 163, 174, 181, 190, 193

[203] ZAJONC, R. Attitudinal effects of "mere exposure". *Journal of personality and social psy-*

chology 9, 2 (1968). 29

[204] ZHANG, K. Z. K., CHEUNG, C. M. K., LEE, M. K. O., AND CHEN, H. Understanding the blog service switching in Hong Kong: An empirical investigation. In *Hawaii International Conference on System Sciences* (jan 2008), p. 269. 150, 158, 159

[205] ZHAO, L. Study on online banking adoption and its predictors. In *International Conference on Multimedia and Information Technology* (2010), vol. 1, pp. 155–158. 158, 173, 183, 187

[206] ZHAO, L., LU, Y., ZHANG, L., AND CHAU, P. Y. K. Assessing the effects of service quality and justice on customer satisfaction and the continuance intention of mobile value-added services: An empirical test of a multidimensional model. *Decision Support Systems 52*, 3 (feb 2012), 645–656. 10, 49, 158, 169, 187

[207] ZHOU, T. The impact of perceived value on user acceptance of mobile commerce. In *International Symposium on Electronic Commerce and Security* (2008), pp. 237–240. 158, 190, 195

[208] ZHOU, T., AND LU, Y. Examining mobile instant messaging user loyalty from the perspectives of network externalities and flow experience. *Computers in Human Behavior 27*, 2 (mar 2011), 883–889. 151, 167, 175, 193

[209] ZHOU, T., AND ZHANG, S. Examining the effect of e-commerce website quality on user satisfaction. In *International Symposium on Electronic Commerce and Security* (may 2009), vol. 1, pp. 418–421. 163, 190, 193

[210] ZHOU, T., ZHANG, S., AND JI, B. Exploring the effect of online banking service quality on users' continuance usage. In *International Conference on E-business and Information System Security* (may 2010), pp. 1–4. 158, 183

[211] ZHOU, X., AND SUN, G. A study of the critical factors that impact users satisfaction in ERP implementations in China. In *International Conference on Information Science and Engineering* (dec 2009), pp. 2824–2826. 173, 183, 187

[212] ZHU, D.-S., TSAI, C.-H., LAN, Y.-L., AND LI, D.-L. A study on the using behavior of depot-logistic information system in Taiwan: An integration of satisfaction theory and technology acceptance theory. *ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 8*, 2 (aug 2012), 347–352. 163, 173, 183, 187, 193

[213] ZVIRAN, M., GLEZER, C., AND AVNI, I. User satisfaction from commercial web sites: The effect of design and use. *Information and Management 43*, 2 (mar 2006), 157–178. 156, 177, 192

List of Figures

List of Tables

Publications

We published parts of the literature reviews, formal concepts and evaluation results of this thesis as scientific publications. In the following, we provide an overview of the relevant publications in chronological order:

1. Koch, S. H., Proynova, R., Paech, B., and Wetter, T. (2014). The perfectly motivated nurse. Journal of Nursing Management 22(8), pp.1054-1064.

2. Proynova, R., and Paech, B. (2013). Factors influencing user feedback on predicted satisfaction with software systems. In 19th International Working Conference on Requirements Engineering: Foundation for Software Quality, pp. 96-111.

3. Proynova, R., and Paech, B. (2012). Do Stakeholders Understand Feature Descriptions? A Live Experiment. In 18th International Working Conference on Requirements Engineering: Foundation for Software Quality, pp. 265–280.

4. Proynova, R., Paech, B., Koch, S. H., Wicht, A., and Wetter, T. (2011). Investigating the Influence of Personal Values on Requirements for Health Care Information Systems. In 3rd international workshop on Software Engineering in Health Care, pp. 48-55.

5. Proynova, R., and Paech, B. (2010). Use of Personal Values in Requirements Engineering-A Research Preview. In 16th International Working Conference on Requirements Engineering: Foundation for Software Quality. pp. 17–22. Essen (Germany).

A List of concepts related to satisfaction

This list shows some detail of the data we gathered in the systematic literature review. We show a short description of each concept related to satisfaction. This description has following elements:

ASMA category The category we assigned to the concept using the ASMA categorization schema 4.1

definition The chosen definition

definition source In order to better understand the concepts and prepare them for further analysis, we needed unambigous definitions. The starting point were definitions provided in the articles studying the concept. If we used one of these definition directly, we note it as "from a primary study" and cite the primary study from which it stems.

In some cases, we made a minor adaptation in the wording to reflect our research context, for example modifying the definition of *adaptability* in [174] from "the degree to which a technology, good, or service can be created, selected, or changed to comply with user preferences." to "The degree to which a *system* can be created, selected, or changed to comply with user preferences." This case is described as "based on a primary study", followed by a reference.

If the primary sources contained definitions which were not adequately describing the concept as used accross studies, or had no definitions at all, we tried using a dictionary definition of the word, either directly cited from a dictionary [147], or slightly reworded analogous to the previous example. We report these cases as "from a dictionary definition" and "based on a dictionary definition" respectively.

Alternatively, if we were aware of scientific research concentrated on this concept, we used a definition from the relevant literature, for example we used a definition of *learnability* provided in a usability measurement textbook [179]. We describe this case as "from a specialized source" followed by a reference to the source.

In cases where this strategy also failed, we created a new definition based on our best understanding of the concept and its use in the primary studies. We report the source in this cases as "our definition".

relationship strength This is the strength we determined in chapter 4.4. It was only calculated for constructs with at least five correlation measurements. The p-value corresponds

to a one-sided test of the estimated correlation coefficient being higher than the center of the next lower strength category.

forest plot Each concept description also includes a forest plot of the correlation measurements in our dataset. It also shows the estimate from a metaanalysis, if conducted.

relationship to satisfaction measured The number of publications which contained a relationship measurement between that concept and satisfaction, used for answering RQ 6. The number of times it was measured as a correlation coefficient was relevant for the metaanalysis for RQ 7. The citations are for the publications in which a relationship was measured (correlation or otherwise).

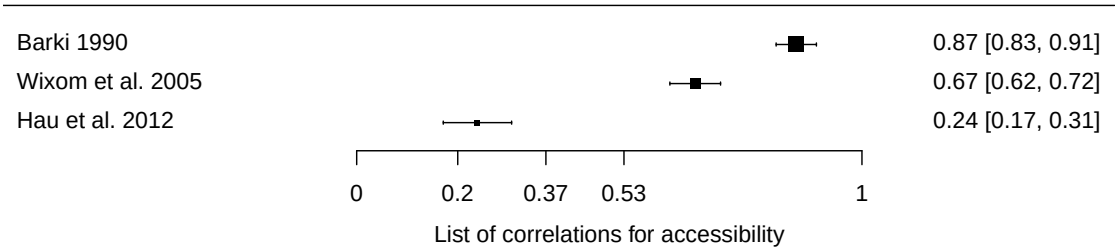
measured for following types of system Lists the types of software system used in the measurements of this concept and its relationship to satisfaction.

A.1 Accessibility

ASMA category system stable properties

definition Refers to the speed of access and availability of the Web site at all times. (from a primary study [37])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [14, 195, 84]

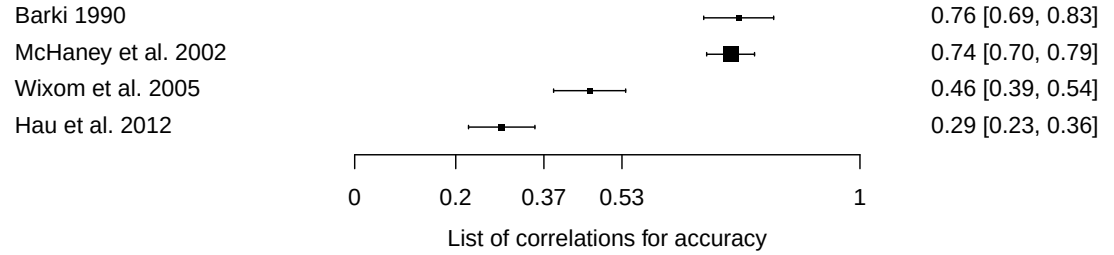
measured for following types of system blog, business information system, e-government, mobile system

A.2 Accuracy

ASMA category information stable properties

definition Information is accurate when it is free from mistakes or errors. (based on a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 6 publications, 4 times as a correlation coefficient

[14, 76, 146, 195, 84, 38]

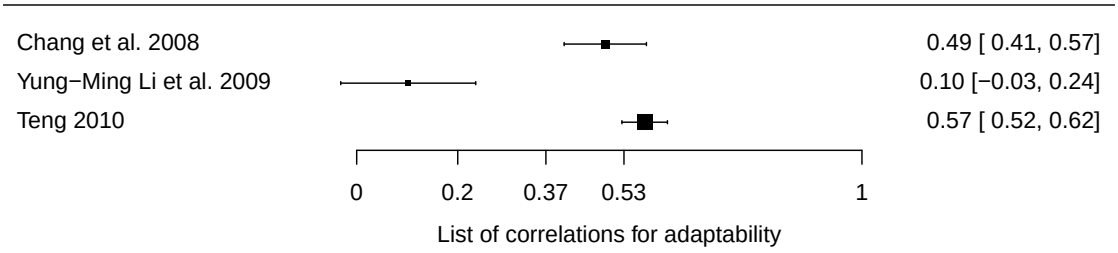
measured for following types of system business information system, e-government, e-learning, mobile system, not system specific

A.3 Adaptability

ASMA category system stable properties

definition The degree to which a system can be created, selected, or changed to comply with user preferences. (based on a primary study [174])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [28, 174, 202]

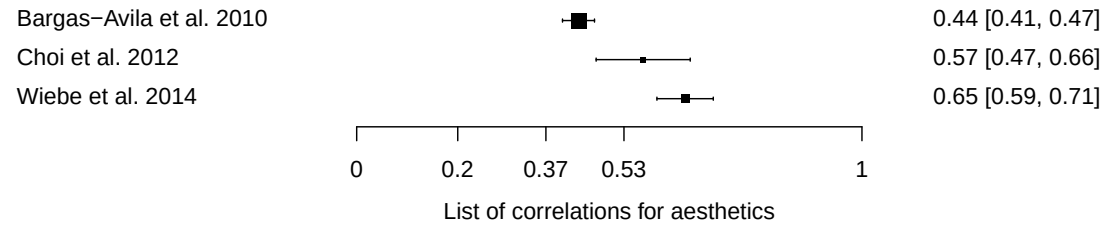
measured for following types of system entertainment, mobile system, unspecified website

A.4 Aesthetics

ASMA category system appraisal

definition A predom- inantly affect-driven evaluative response to the visual Gestalt of an object. (from a primary study [83])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 5 publications, 3 times as a correlation coefficient [13, 40, 91, 193, 127]

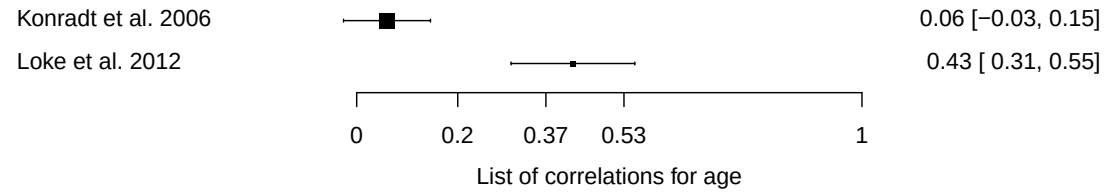
measured for following types of system e-government, entertainment, mobile system, not system specific, unspecified website

A.5 Age

ASMA category user stable properties

definition The amount of time during which the user has lived (based on a primary study [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [141, 116]

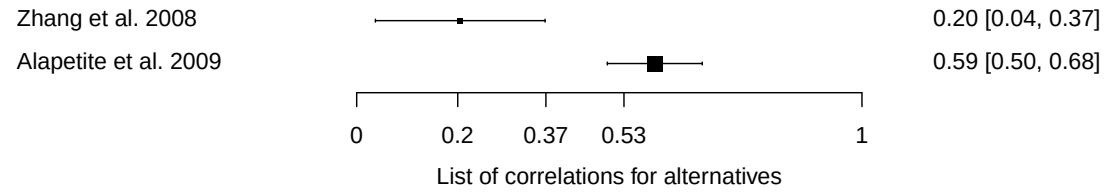
measured for following types of system business information system, online banking

A.6 Alternatives

ASMA category context stable properties

definition The availability of other approaches to execute the same task (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [6, 204]

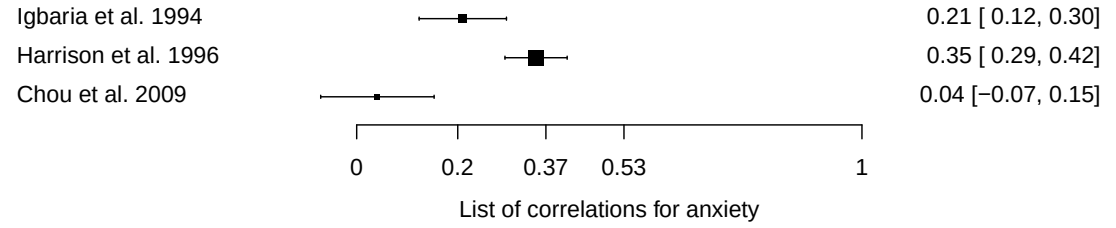
measured for following types of system blog, other

A.7 Anxiety

ASMA category user mutable properties

definition the tendency of an individual to be uneasy, apprehensive, and/or phobic towards current o r future use of computers in general. (from a primary study [96])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 3 times as a correlation coefficient [41, 79, 96, 105]

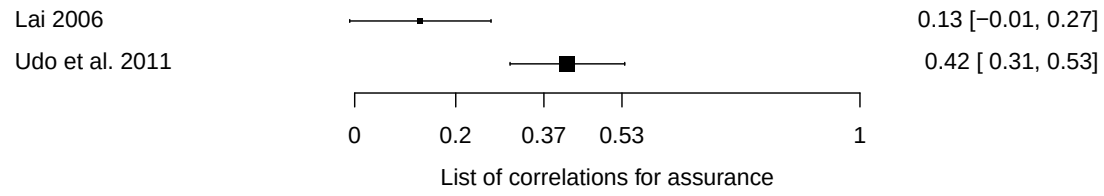
measured for following types of system business information system, mobile system, not system specific

A.8 Assurance

ASMA category system stable properties

definition ability of a system to reduce uncertainties of users (based on a primary study [121])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [97, 121]

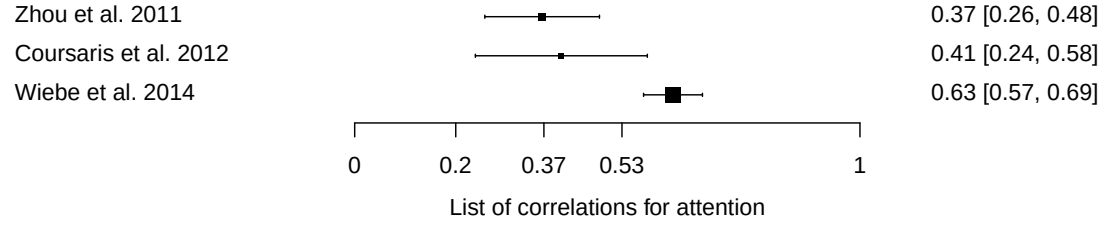
measured for following types of system e-commerce, e-government, e-learning, online banking

A.9 Attention

ASMA category user mutable properties

definition The proportion of time in which the user remains focused on the system and not distracted. (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [46, 193, 208]

measured for following types of system entertainment, mobile system, other, telecommunication network

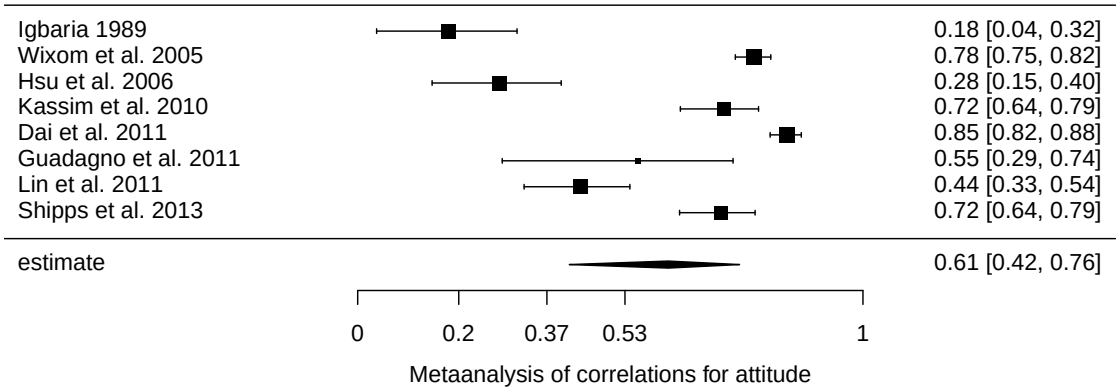
A.10 Attitude

ASMA category user mutable properties

definition attitudes indicate an individual's reaction to or evaluation of an object on a llke-dlslike or favorable-unfavorable continuum (from a primary study [95])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.107$

heterogeneity measures $I^2 = 96.76\%$, $\tau^2 = 0.14$



relationship to satisfaction measured in 11 publications, 8 times as a correlation coefficient [5, 48, 75, 86, 92, 95, 107, 137, 160, 195, 169]

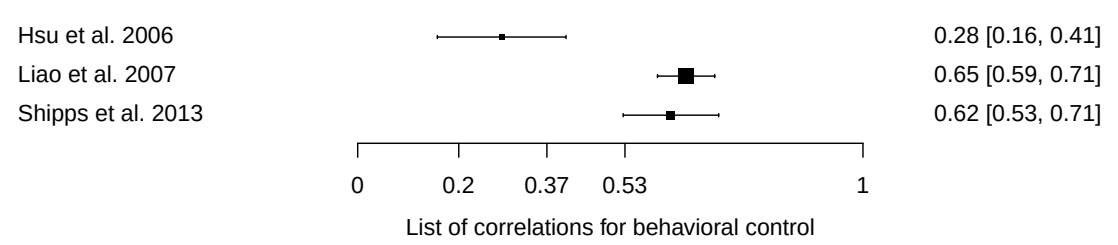
measured for following types of system business information system, e-commerce, e-government, e-learning, entertainment, not system specific, online community, other

A.11 Behavioral control

ASMA category user mutable properties

definition Behavioral control reflects one's perceptions of the availability of resources or opportunities necessary for performing a behavior (from a primary study [130])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 3 times as a correlation coefficient [92, 130, 139, 169]

measured for following types of system business information system, e-commerce, e-learning, online community

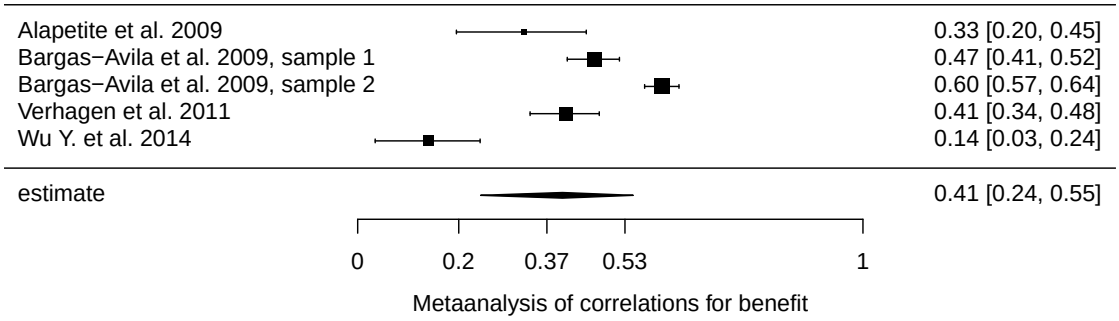
A.12 Benefit

ASMA category context mutable properties

definition the user's net gain from acquiring or using a software system, which is clearly an extrinsic value related to commercial activities. (from a primary study [185])

effect strength for satisfaction relationship Medium. Difference from lower strength is not significant, $p = 0.100$

heterogeneity measures $I^2 = 96.04\%$, $\tau^2 = 0.04$



relationship to satisfaction measured in 5 publications, 5 times as a correlation coefficient [6, 12, 182, 197, 185]

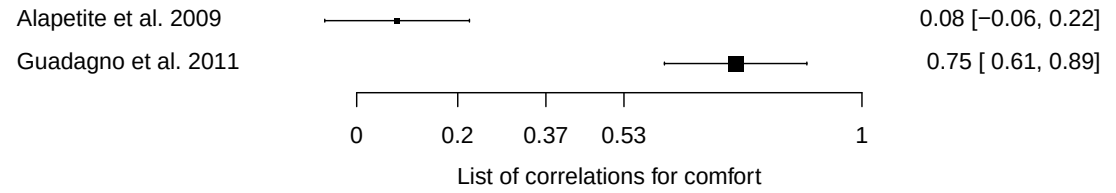
measured for following types of system business information system, e-commerce, entertainment, not system specific, online community, other

A.13 Comfort

ASMA category system stable properties

definition a feeling of relief or encouragement (from a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 2 times as a correlation coefficient [6, 75, 33]

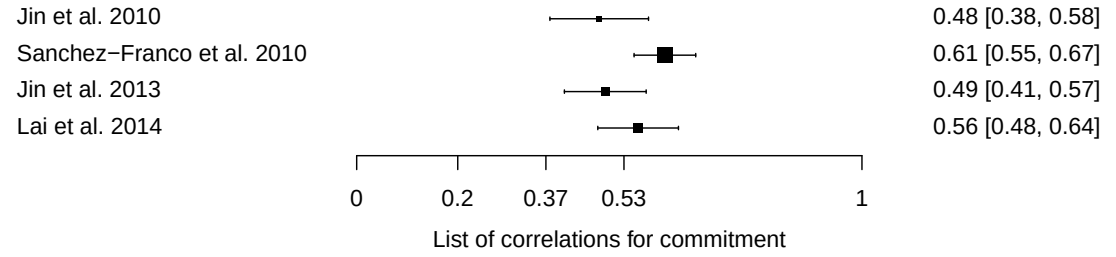
measured for following types of system entertainment, online banking, other

A.14 Commitment

ASMA category user stable properties

definition A persistent desire to keep a valuable relationship (from a primary study [102])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 4 times as a correlation coefficient [103, 102, 162, 120]

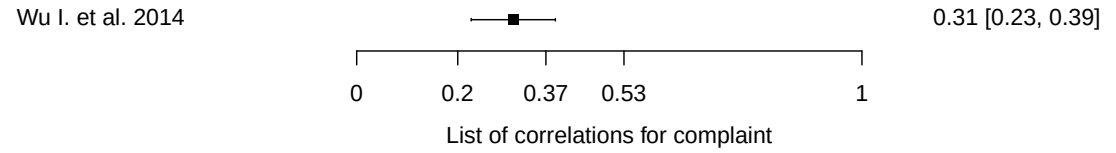
measured for following types of system entertainment, mobile system, online community, other

A.15 Complaint

ASMA category user activity

definition expression of grief, pain, or dissatisfaction (from a primary study [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [191, 196]

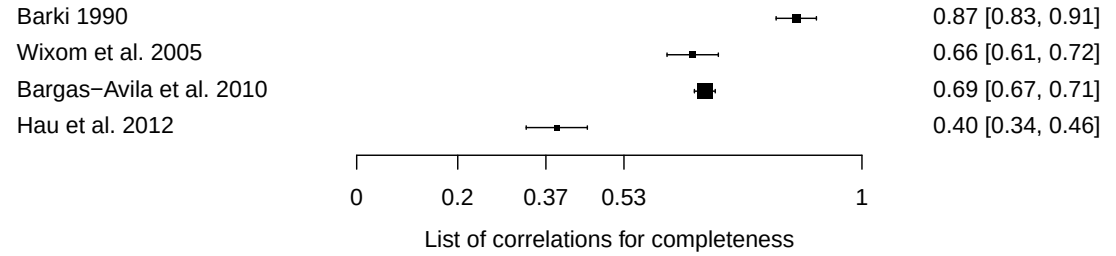
measured for following types of system e-commerce

A.16 Completeness

ASMA category information stable properties

definition Information which has all necessary parts, not lacking anything (from a primary study [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 5 publications, 4 times as a correlation coefficient [13, 14, 195, 84, 38]

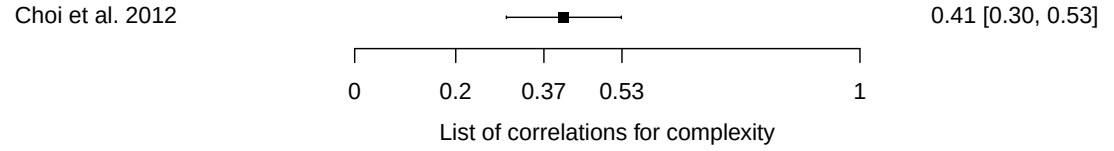
measured for following types of system business information system, e-government, e-learning, mobile system, other

A.17 Complexity

ASMA category system stable properties

definition Component complexity is a function of the number of distinct information cues that must be processed in the performance of a task (from a primary study [40])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [40, 198]

measured for following types of system mobile system, not system specific

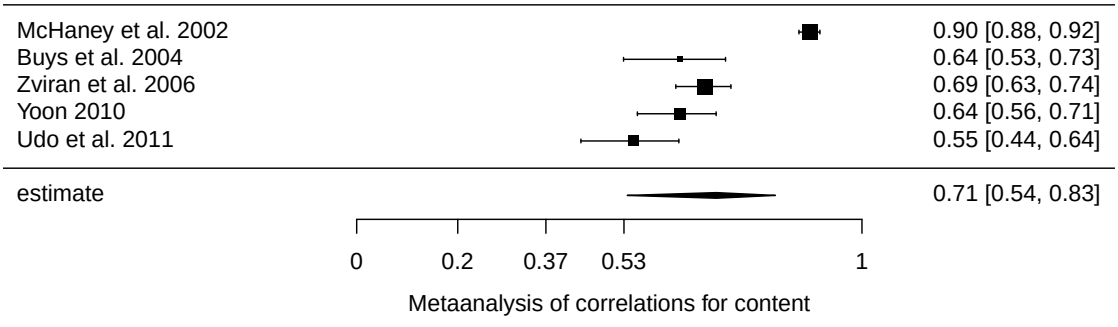
A.18 Content

ASMA category information stable properties

definition the inherent value and usefulness of the information provided by the system (from a primary study [115])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.008$

heterogeneity measures $I^2 = 96.29\%$, $\tau^2 = 0.11$



relationship to satisfaction measured in 5 publications, 5 times as a correlation coefficient [25, 115, 201, 146, 213]

measured for following types of system business information system, e-commerce, e-learning,

mobile system, online banking

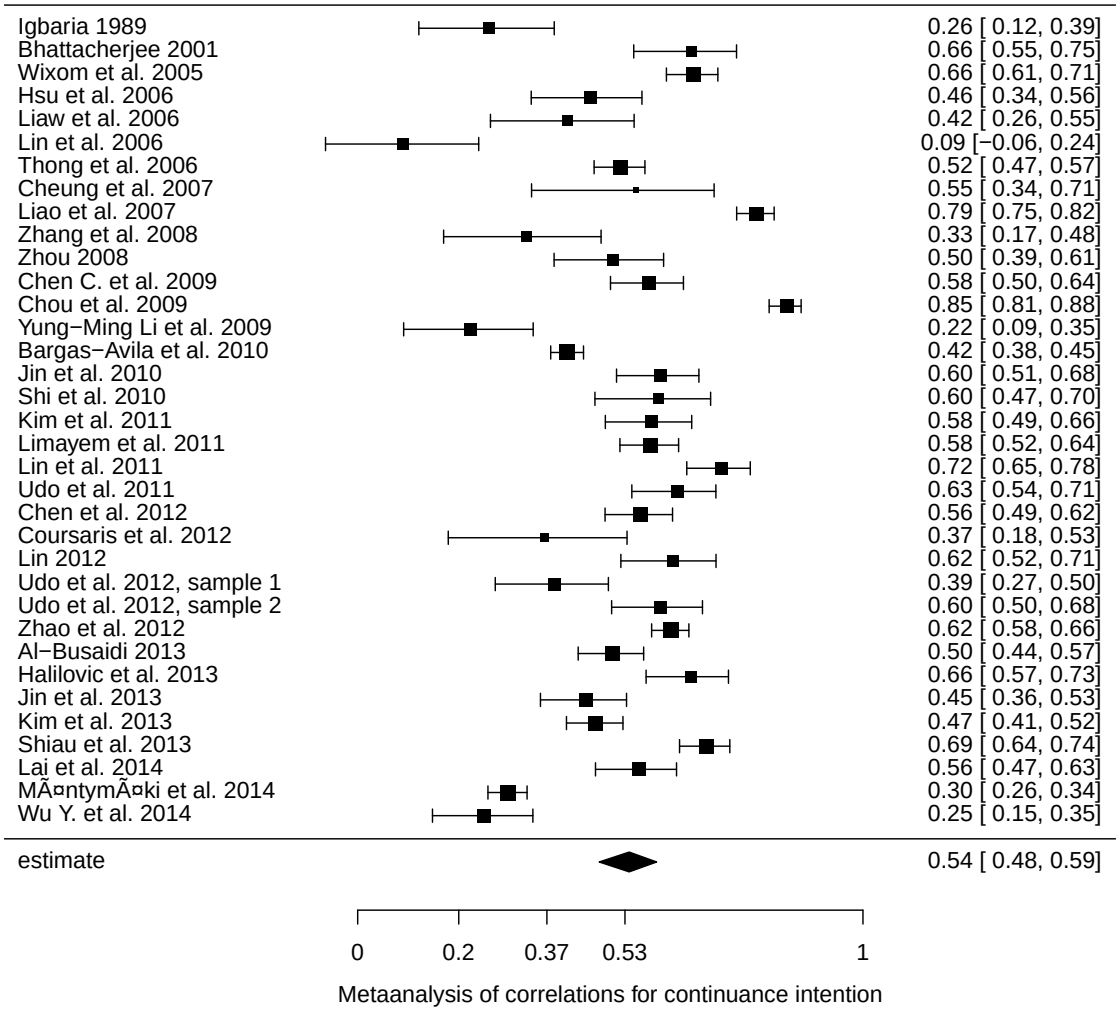
A.19 Continuance intention

ASMA category user activity

definition The user's decision to continue using the system over a long period of time (from a primary study [20])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p < 0.001$

heterogeneity measures $I^2 = 95.59\%$, $\tau^2 = 0.05$



relationship to satisfaction measured in 50 publications, 35 times as a correlation coefficient

[4, 13, 20, 21, 32, 34, 36, 41, 44, 46, 74, 77, 86, 92, 95, 101, 103, 102, 105, 115, 117, 130, 131, 133, 135, 136, 137, 138, 167, 172, 176, 180, 181, 189, 197, 202, 204, 205, 207, 210, 144, 132, 108, 195, 35, 168, 120, 33, 109, 206]

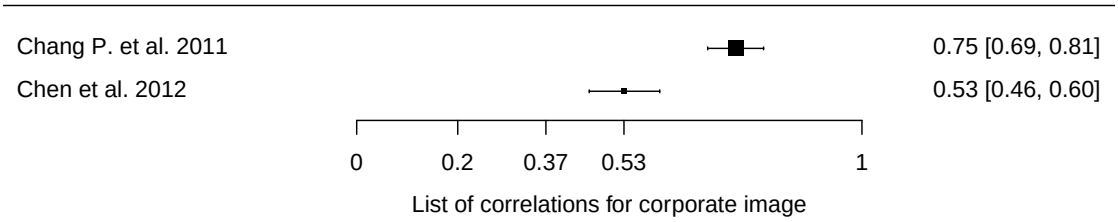
measured for following types of system blog, business information system, e-commerce, e-government, e-learning, entertainment, mobile services, mobile system, online banking, online community, other

A.20 Corporate image

ASMA category context stable properties

definition the consequence of the comparison between what the company promises and what it eventually fulfils. Thus, reputation would show how honest the company is and how much it cares for its environment (from a primary study [27])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 2 times as a correlation coefficient [5, 30, 34, 128]

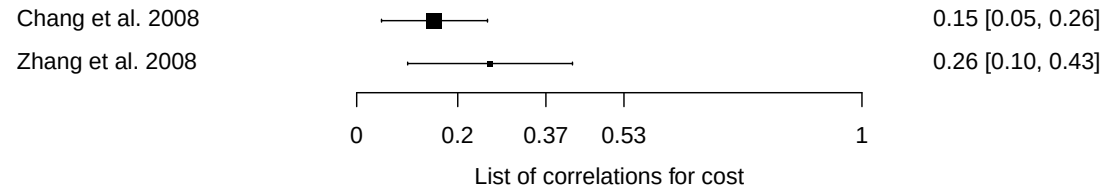
measured for following types of system business information system, e-commerce, e-learning, online banking, telecommunication network

A.21 Cost

ASMA category context stable properties

definition The amount of money the user must expend in order to use the system. (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [28, 204]

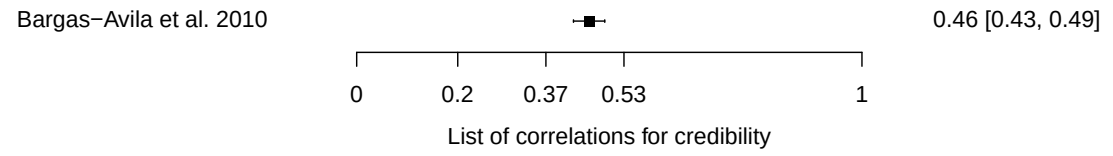
measured for following types of system blog, not system specific, online community, unspecified website

A.22 Credibility

ASMA category system stable properties

definition Design credibility is defined as a holistic concept that covers an online user's perception of safety, reliability, security, and privacy during the navigation of the website (from a primary study [74])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [13, 117]

measured for following types of system e-commerce, e-government, other

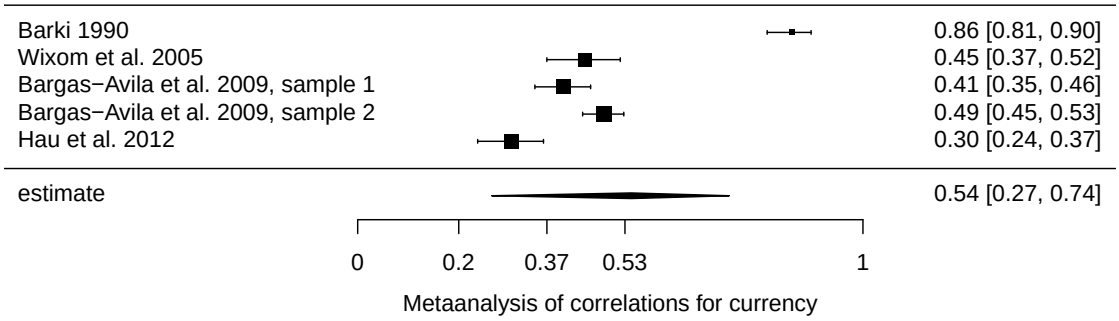
A.23 Currency

ASMA category information mutable properties

definition Information which is up-to-date (based on a primary study [55])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.213$

heterogeneity measures $I^2 = 98.95\%$, $\tau^2 = 0.14$



relationship to satisfaction measured in 5 publications, 5 times as a correlation coefficient [12, 14, 195, 84, 38]

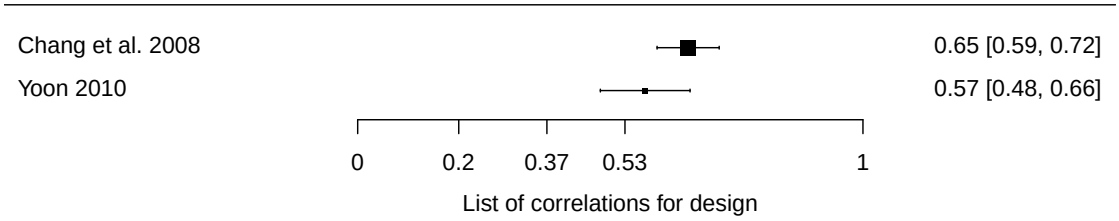
measured for following types of system business information system, e-learning, mobile system, not system specific

A.24 Design

ASMA category system appraisal

definition the overall image or personality that the system provider projects to users using inputs such as text, style, graphics, colors, logos, and slogans or themes (from a primary study [28])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 2 times as a correlation coefficient [28, 201, 126]

measured for following types of system business information system, mobile system, online banking, unspecified website

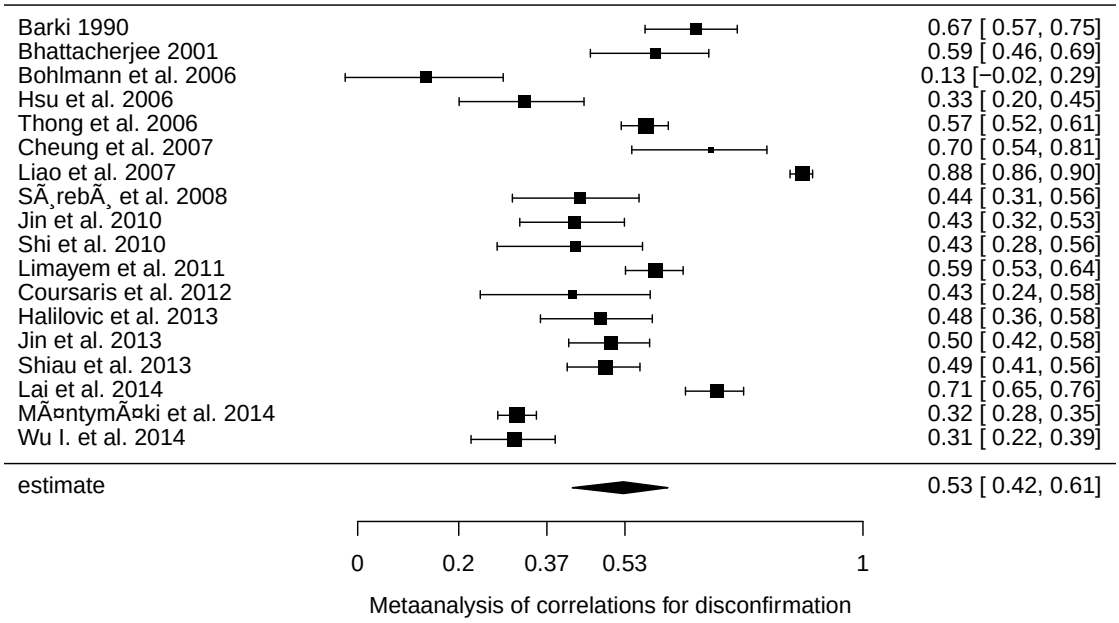
A.25 Disconfirmation

ASMA category user mutable properties

definition The evaluation of the empirical gap between expectation and the results of actual usage. (from a primary study [102])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.032$

heterogeneity measures $I^2 = 96.27\%$, $\tau^2 = 0.08$



relationship to satisfaction measured in 22 publications, 18 times as a correlation coefficient [14, 20, 22, 36, 46, 47, 77, 92, 103, 102, 130, 133, 142, 167, 171, 172, 176, 144, 35, 168, 120, 196]

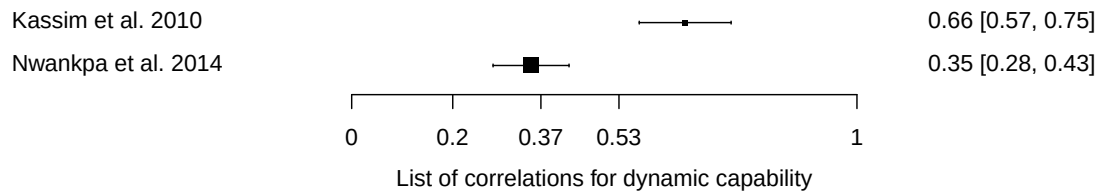
measured for following types of system blog, business information system, e-commerce, e-learning, entertainment, mobile services, mobile system, online banking, online community, other, unspecified website

A.26 Dynamic capability

ASMA category context stable properties

definition the ability to integrate, build, and reconfigure internal and external competencies to address rapidly-changing environments (from a primary study [107])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [107, 150]

measured for following types of system business information system, e-government

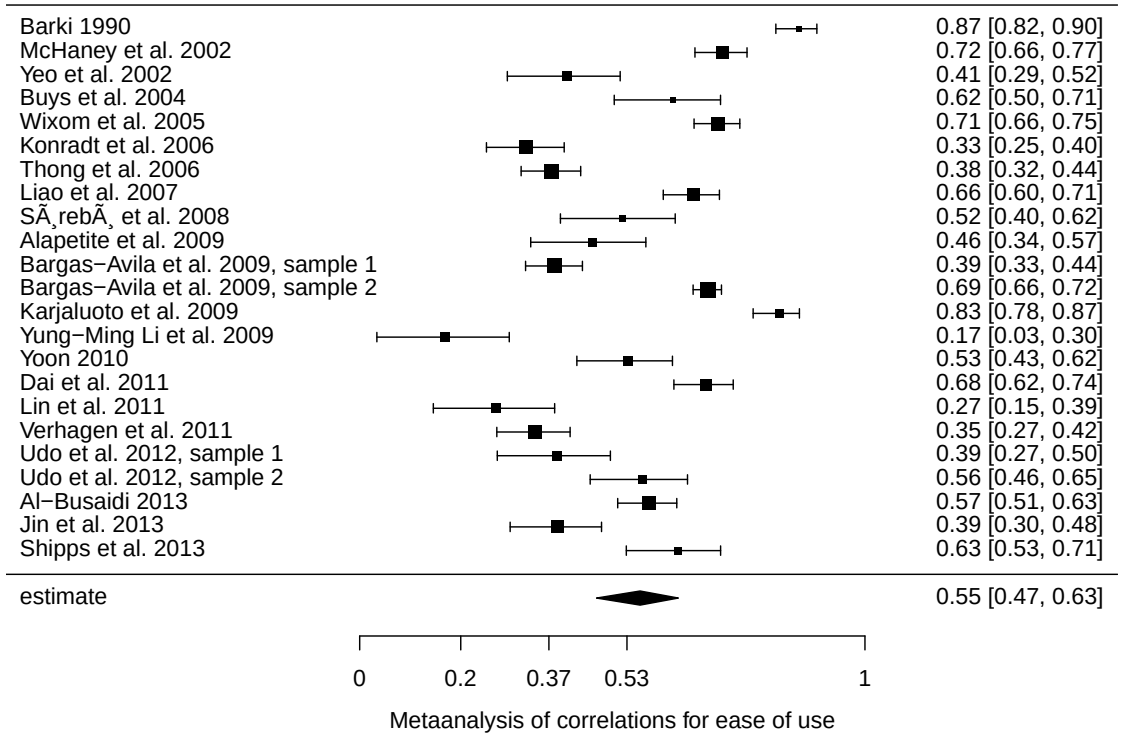
A.27 Ease of use

ASMA category system appraisal

definition the degree to which an individual believes that using a particular system would be free from physical and mental effort (from a primary study [5])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.001$

heterogeneity measures $I^2 = 96.63\%$, $\tau^2 = 0.08$



relationship to satisfaction measured in 30 publications, 23 times as a correlation coefficient [4, 5, 6, 12, 13, 14, 25, 26, 48, 74, 86, 102, 130, 137, 139, 154, 171, 176, 181, 200, 201, 202, 209, 212, 146, 185, 195, 169, 106, 116]

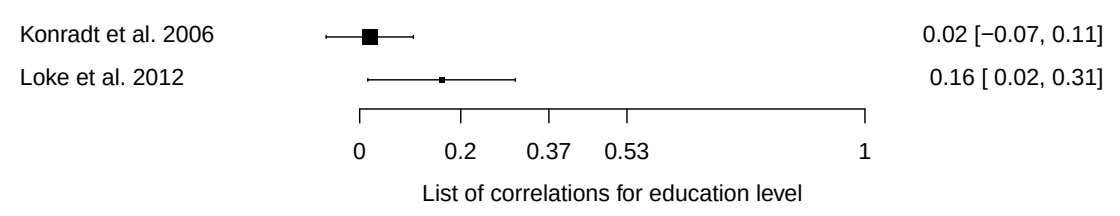
measured for following types of system business information system, e-commerce, e-government, e-learning, entertainment, mobile services, mobile system, not system specific, online banking, online community, other, unspecified website

A.28 Education level

ASMA category user stable properties

definition The degree of formal education reached by the user (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [141, 116]

measured for following types of system business information system, online banking

A.29 Effectiveness

ASMA category system activity

definition the accuracy and completeness with which users achieve certain goals. (from a primary study [68])

effect strength for satisfaction relationship No metaanalysis conducted

relationship to satisfaction measured in 2 publications, 0 times as a correlation coefficient [46, 76]

measured for following types of system business information system, mobile system, not system specific

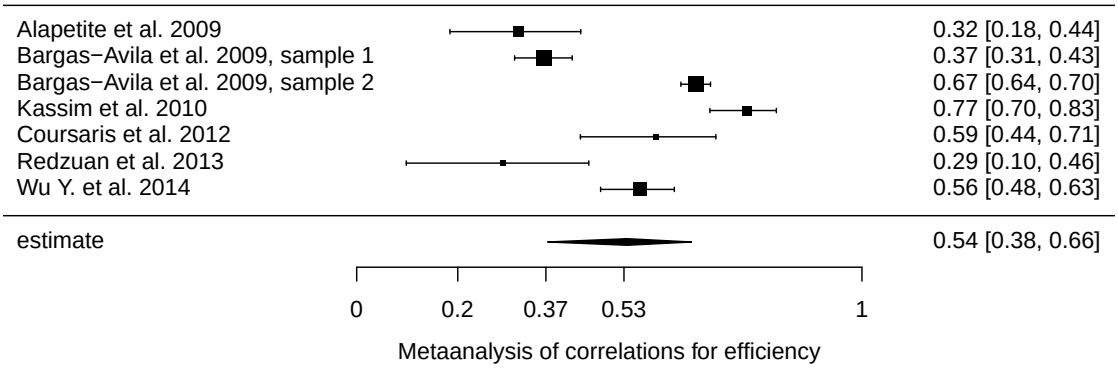
A.30 Efficiency

ASMA category system activity

definition the level of resources consumed in performing tasks (from a primary study [46])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.159$

heterogeneity measures $I^2 = 96.08\%$, $\tau^2 = 0.07$



relationship to satisfaction measured in 6 publications, 7 times as a correlation coefficient [6, 12, 46, 107, 159, 197]

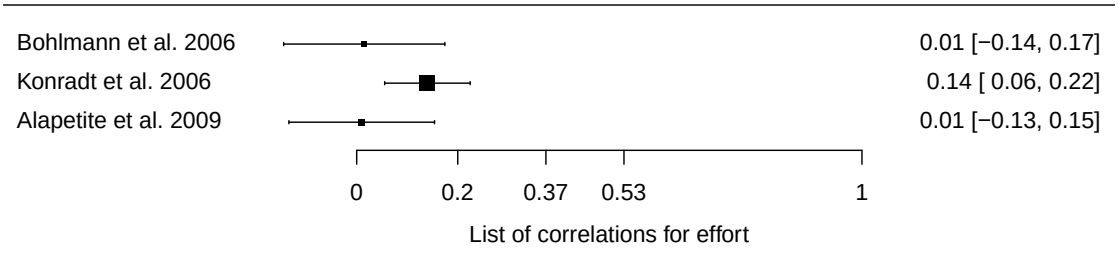
measured for following types of system business information system, e-government, mobile system, not system specific, online banking, online community, other

A.31 Effort

ASMA category user activity

definition conscious exertion of power, work done by the mind or body (from a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 3 times as a correlation coefficient [6, 22, 76, 116]

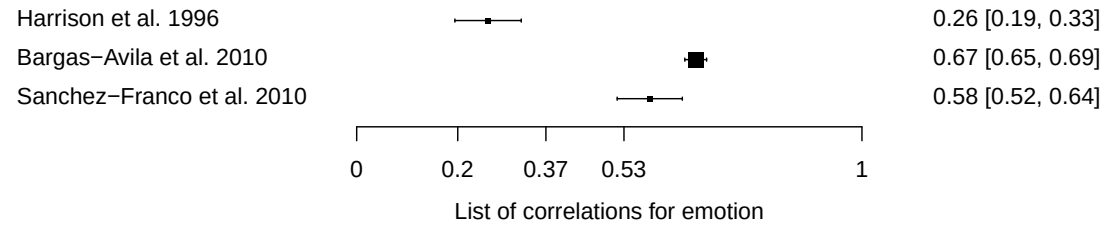
measured for following types of system business information system, e-learning, other

A.32 Emotion

ASMA category user mutable properties

definition a conscious mental reaction (as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body (from a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 3 times as a correlation coefficient [13, 79, 191, 162]

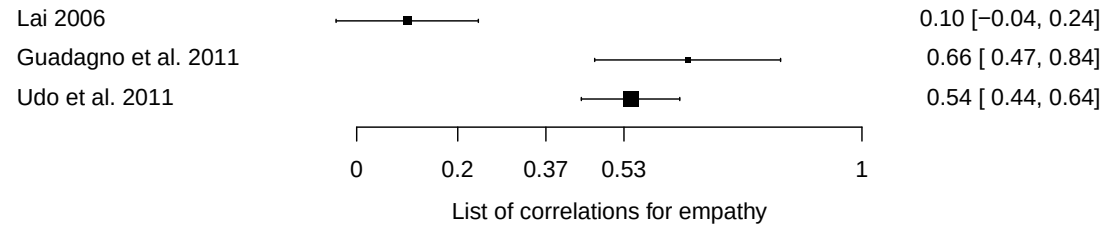
measured for following types of system e-commerce, e-government, e-learning, entertainment, not system specific

A.33 Empathy

ASMA category system stable properties

definition Degree to which a system understands and reflects individual states and desires of users (from a primary study [97])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [75, 97, 121]

measured for following types of system e-commerce, e-government, e-learning, entertainment, online banking

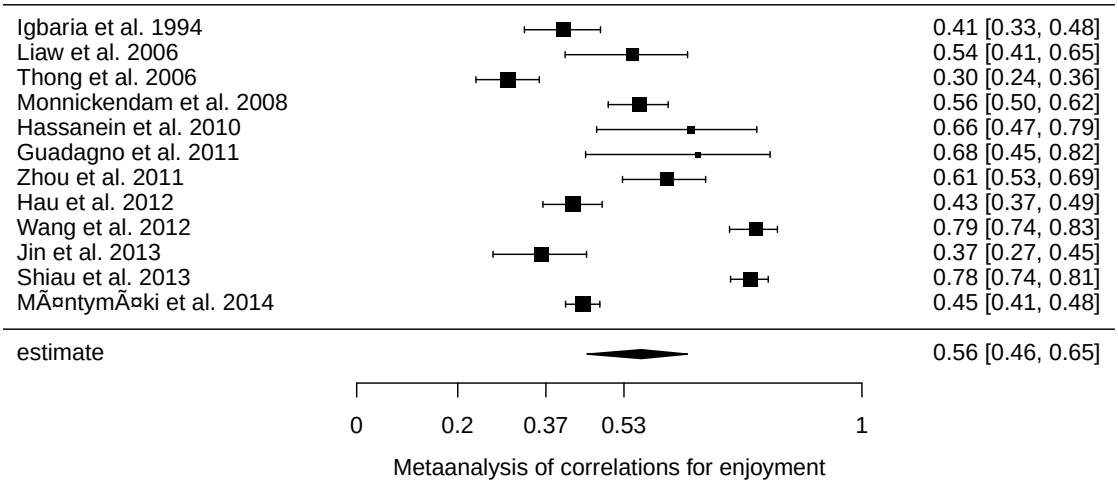
A.34 Enjoyment

ASMA category user mutable properties

definition The degree to which users believe that using a specific system is fun and rewarding (from a primary study [148])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.009$

heterogeneity measures $I^2 = 96.5\%$, $\tau^2 = 0.06$



relationship to satisfaction measured in 14 publications, 12 times as a correlation coefficient [5, 75, 82, 96, 102, 105, 131, 148, 176, 208, 144, 188, 168, 84]

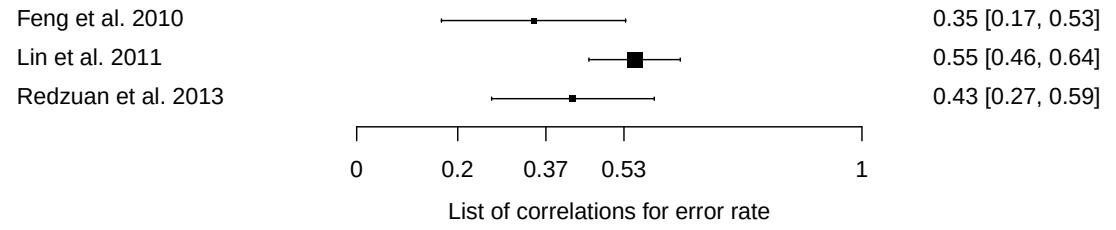
measured for following types of system blog, e-learning, entertainment, mobile services, mobile system, not system specific, other, telecommunication network

A.35 Error rate

ASMA category system activity

definition The proportion of incorrect actions taken by the system (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [61, 137, 159]

measured for following types of system e-learning, other

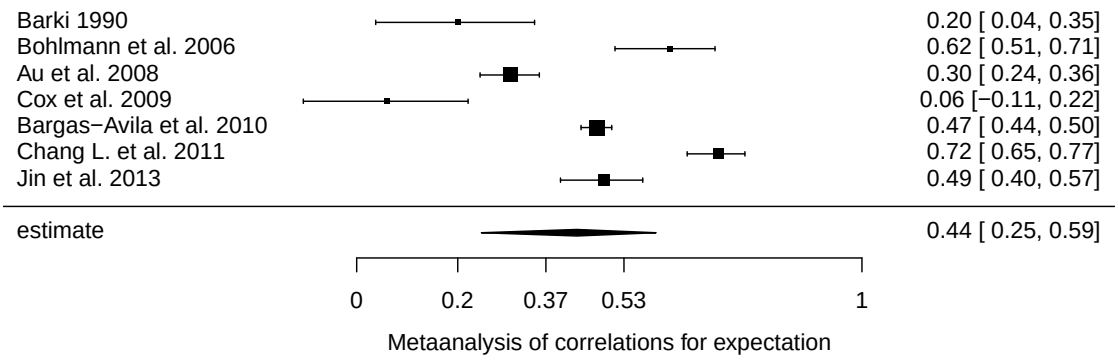
A.36 Expectation

ASMA category user mutable properties

definition Expectations are beliefs or subjective predictions about a system's attributes or performance at some time in the future, or the likelihood that a system is associated with certain attributes, benefits, and outcomes, which are oriented toward the future and are relatively changeable (from a primary study [142])

effect strength for satisfaction relationship Medium. Difference from lower strength is not significant, $p = 0.104$

heterogeneity measures $I^2 = 97.41\%$, $\tau^2 = 0.08$



relationship to satisfaction measured in 9 publications, 7 times as a correlation coefficient [10, 13, 14, 22, 29, 47, 102, 128, 139]

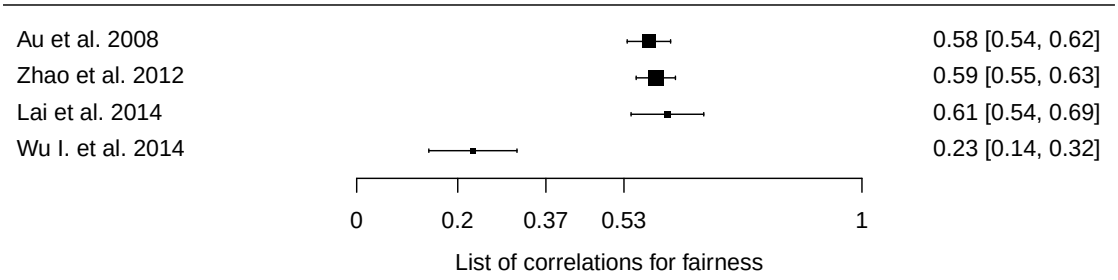
measured for following types of system blog, business information system, e-government, e-learning, mobile system, not system specific, online banking, other, unspecified website

A.37 Fairness

ASMA category context mutable properties

definition equity, rightness or deservingness comparison to other entities, whether real or imaginary, individual or collective, person or nonperson. Also includes justice. (based on a primary study [153])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 4 times as a correlation coefficient [10, 120, 196, 206]

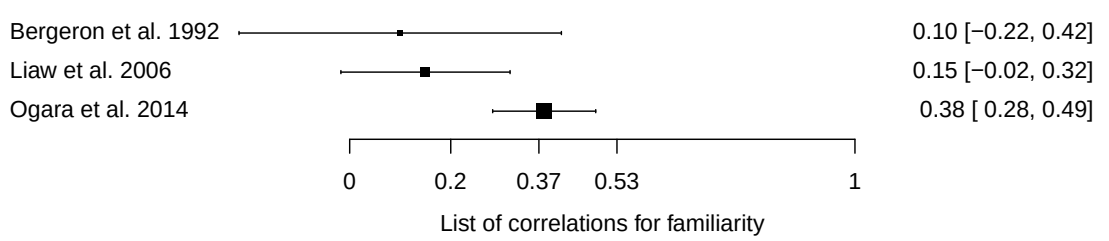
measured for following types of system business information system, e-commerce, mobile system, other

A.38 Familiarity

ASMA category system stable properties

definition The degree to which the user is acquainted with the system (based on a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [18, 131, 151]

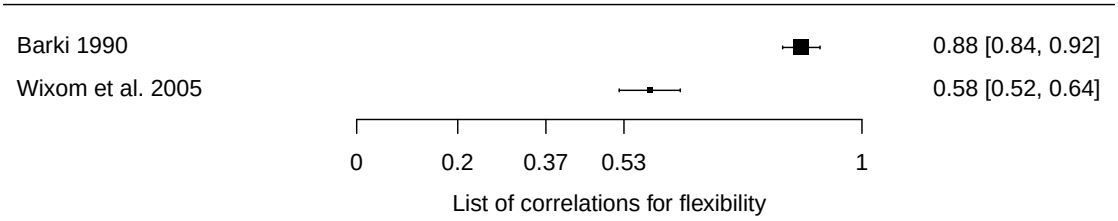
measured for following types of system e-commerce, not system specific, other, telecommunication network

A.39 Flexibility

ASMA category system stable properties

definition A system which is able to change or to do different things (from a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [14, 195]

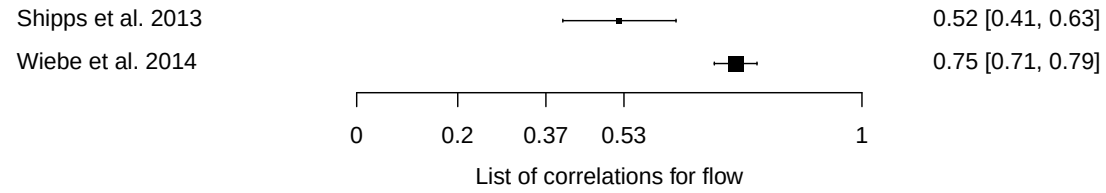
measured for following types of system business information system

A.40 Flow

ASMA category user mutable properties

definition complete immersion in the task (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [193, 169]

measured for following types of system entertainment, online community

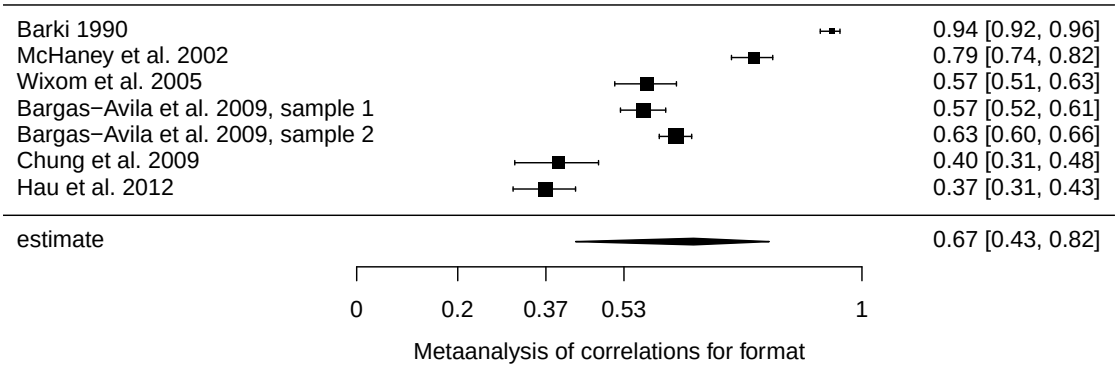
A.41 Format

ASMA category information mutable properties

definition the form, design, or arrangement of the information presented by the system (based on a dictionary definition [147])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.107$

heterogeneity measures $I^2 = 99.18\%$, $\tau^2 = 0.21$



relationship to satisfaction measured in 6 publications, 7 times as a correlation coefficient [12, 14, 42, 146, 195, 84]

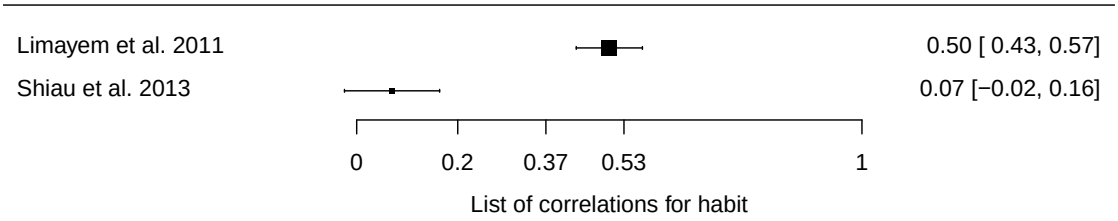
measured for following types of system business information system, mobile system, other

A.42 Habit

ASMA category user activity

definition The user uses the system in frequent, regular intervals (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 2 times as a correlation coefficient [105, 133, 168]

measured for following types of system blog, e-learning, mobile system

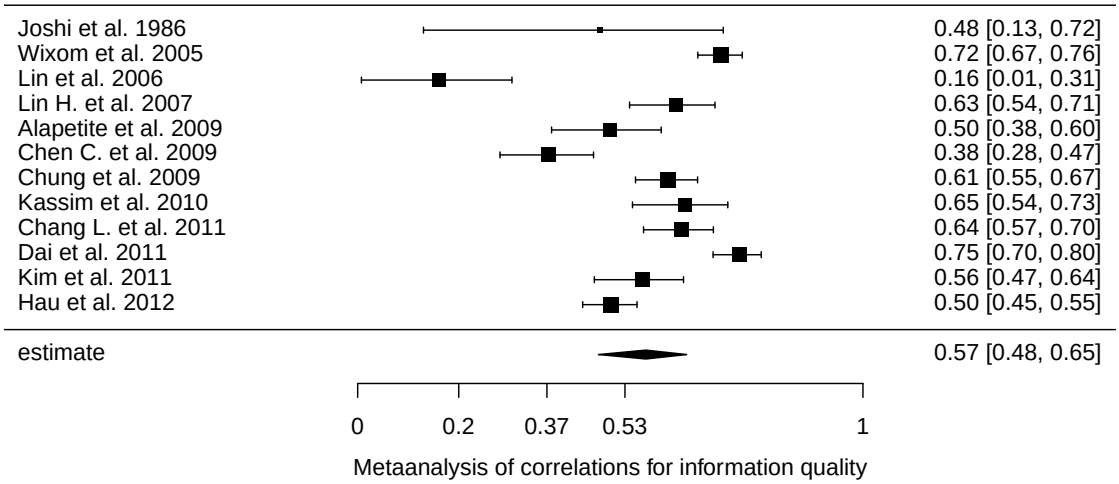
A.43 Information quality

ASMA category information appraisal

definition the technical quality of reports and screens generated by the information systems (from a primary study [104])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.001$

heterogeneity measures $I^2 = 92.85\%$, $\tau^2 = 0.05$



relationship to satisfaction measured in 20 publications, 12 times as a correlation coefficient [6, 32, 31, 42, 48, 101, 104, 105, 107, 135, 134, 182, 191, 205, 211, 212, 108, 195, 84, 126]

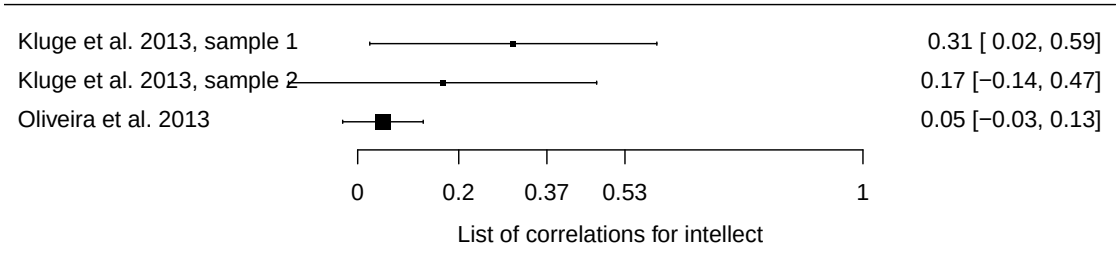
measured for following types of system blog, business information system, e-commerce, e-government, e-learning, mobile system, not system specific, online banking, online community, other, unspecified website

A.44 Intellect

ASMA category user stable properties

definition the ability to learn or understand things or to deal with new or difficult situations (from a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 3 times as a correlation coefficient [113, 152]

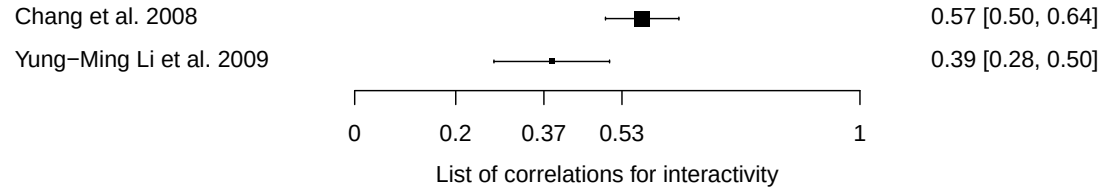
measured for following types of system mobile system, other

A.45 Interactivity

ASMA category system stable properties

definition the availability and effectiveness of user support tools on the system interface and the degree to which the system facilitates two-way communication with customers (based on a primary study [28])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 2 times as a correlation coefficient [28, 142, 202]

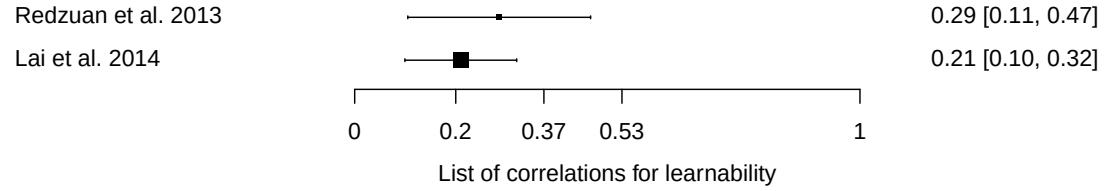
measured for following types of system blog, e-commerce, e-government, mobile system, unspecified website

A.46 Learnability

ASMA category system stable properties

definition the capability of a software product to enable the user to learn how to use it (from a specialized source [179])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [159, 120]

measured for following types of system business information system, online banking, other

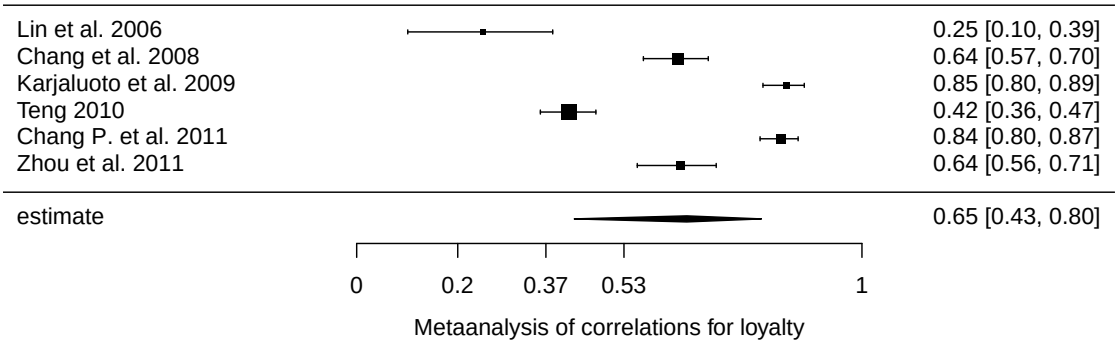
A.47 Loyalty

ASMA category user stable properties

definition the extent to which something can be learned efficiently. (from a primary study [27])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.107$

heterogeneity measures $I^2 = 97.89\%$, $\tau^2 = 0.16$



relationship to satisfaction measured in 8 publications, 6 times as a correlation coefficient [27, 28, 30, 128, 135, 174, 208, 106]

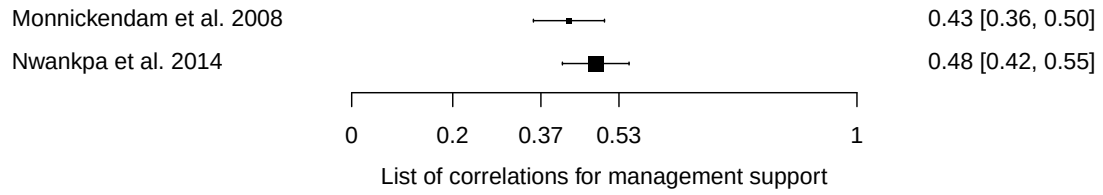
measured for following types of system business information system, e-commerce, entertainment, online banking, online community, telecommunication network, unspecified website

A.48 Management support

ASMA category context stable properties

definition The degree to which the organisation's management supports development of the system (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 2 times as a correlation coefficient [21, 148, 150]

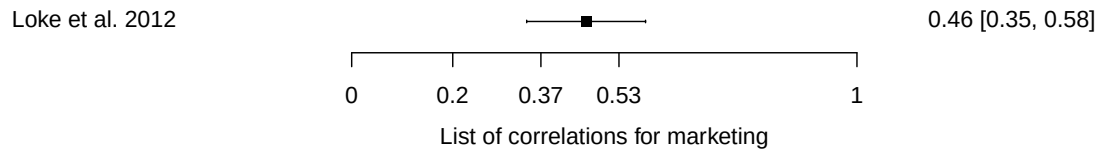
measured for following types of system business information system, not system specific

A.49 Marketing

ASMA category context stable properties

definition Effort made to make the system attractive to potential users (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [141, 199]

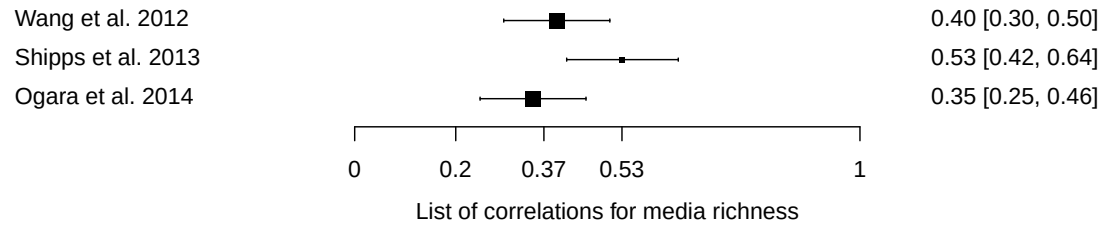
measured for following types of system online banking, telecommunication network

A.50 Media richness

ASMA category information stable properties

definition the amount of information a medium can convey to change the receiver's understanding (based on a primary study [169])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 3 times as a correlation coefficient [151, 188, 169]

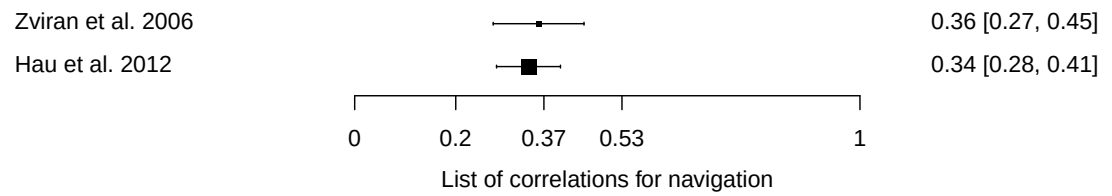
measured for following types of system online community, telecommunication network

A.51 Navigation

ASMA category system stable properties

definition The structure of a system interface as exposed to the user (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [213, 84]

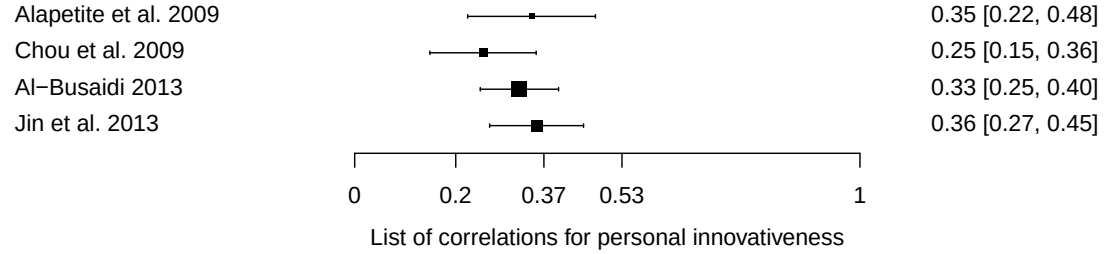
measured for following types of system business information system, e-commerce, mobile system, other, unspecified website

A.52 Personal innovativeness

ASMA category user stable properties

definition the tendency to experiment with and to adopt new information technologies independently of the experience of others (from a primary study [4])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 4 times as a correlation coefficient [4, 6, 102]

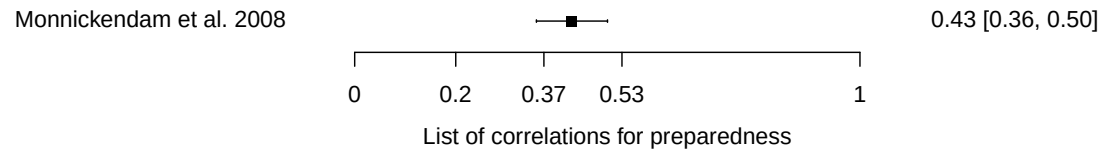
measured for following types of system business information system, e-learning, mobile system, other

A.53 Preparedness

ASMA category user mutable properties

definition users having the appropriate skills to operate the new system. (from a primary study [148])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [136, 148]

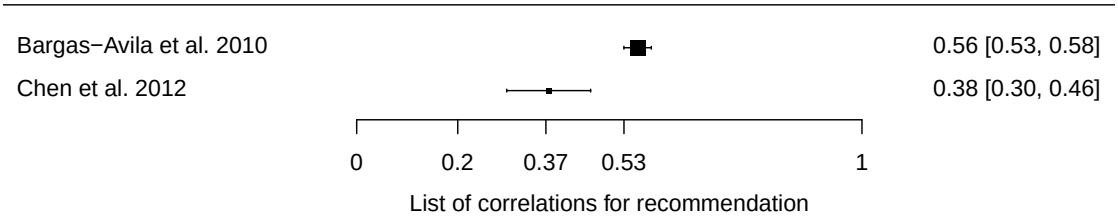
measured for following types of system e-government, not system specific

A.54 Recommendation

ASMA category user activity

definition Intention to recommend the system to a third person (based on a primary study [97])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 2 times as a correlation coefficient [13, 34, 191]

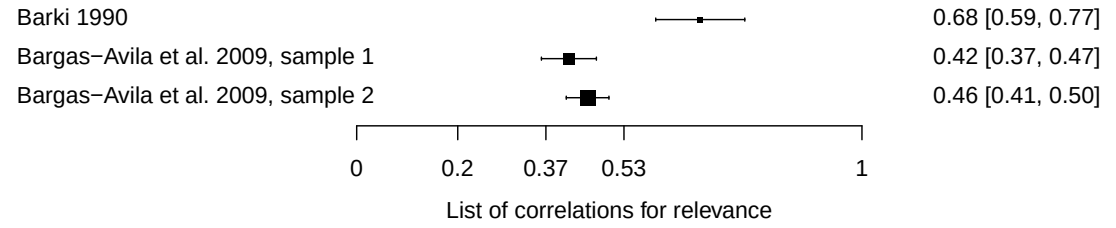
measured for following types of system e-commerce, e-government

A.55 Relevance

ASMA category information mutable properties

definition relating to the task in an appropriate way (based on a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 3 times as a correlation coefficient [12, 14]

measured for following types of system business information system

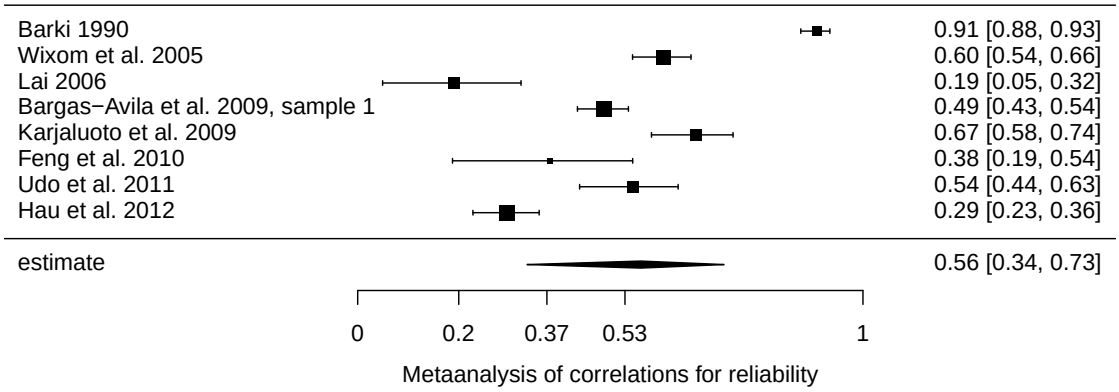
A.56 Reliability

ASMA category system stable properties

definition the ability to perform the promised service dependably and accurately. (from a primary study [180])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.168$

heterogeneity measures $I^2 = 98.18\%$, $\tau^2 = 0.16$



relationship to satisfaction measured in 10 publications, 8 times as a correlation coefficient [8, 12, 14, 61, 97, 101, 121, 195, 106, 84]

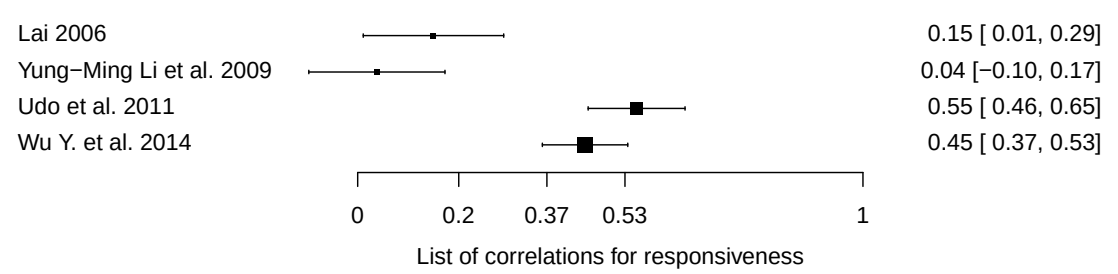
measured for following types of system blog, business information system, e-commerce, e-government, e-learning, mobile system, online banking, other, unspecified website

A.57 Responsiveness

ASMA category system stable properties

definition Degree to which a system voluntarily and quickly responds to demands of consumers (based on a primary study [97])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 5 publications, 4 times as a correlation coefficient [97, 101, 121, 197, 202]

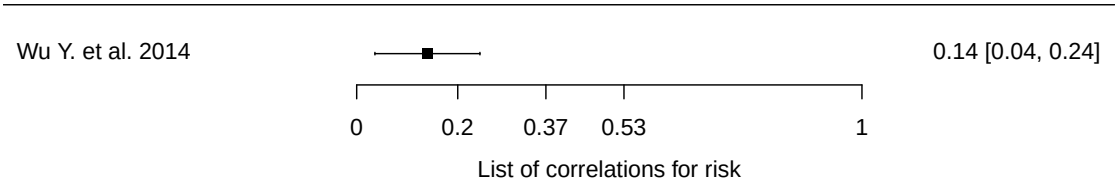
measured for following types of system e-commerce, e-government, e-learning, mobile system, online banking, online community

A.58 Risk

ASMA category context mutable properties

definition the user's subjective belief of suffering a loss in pursuit of a desired outcome (from a primary study [74])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [74, 197]

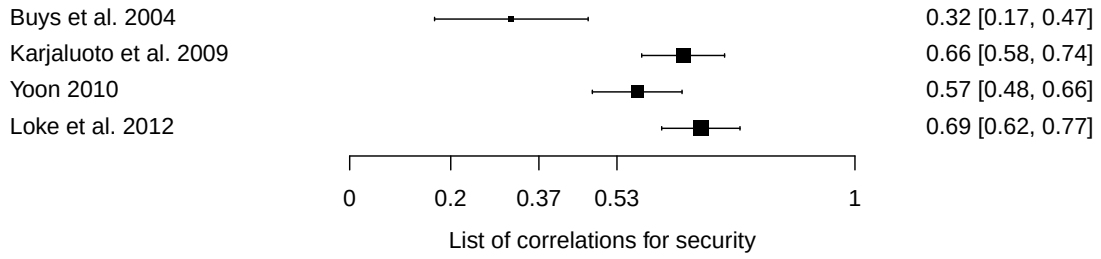
measured for following types of system e-commerce, online community

A.59 Security

ASMA category system stable properties

definition Users' recognition that the adoption of a certain system is risk free. (from a primary study [187])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 8 publications, 4 times as a correlation coefficient [8, 25, 101, 141, 173, 201, 33, 106]

measured for following types of system e-commerce, e-government, mobile system, online banking, online community, unspecified website

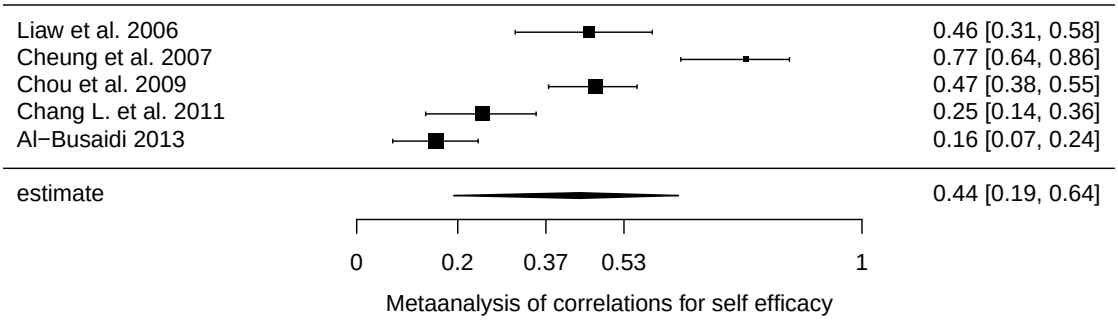
A.60 Self efficacy

ASMA category user mutable properties

definition Users' judgments of their capabilities to organise and execute courses of action required to attain designated types of performances (from a primary study [4])

effect strength for satisfaction relationship Medium. Difference from lower strength is not significant, $p = 0.160$

heterogeneity measures $I^2 = 95.65\%$, $\tau^2 = 0.09$



relationship to satisfaction measured in 5 publications, 5 times as a correlation coefficient [4, 21, 41, 86, 131]

measured for following types of system business information system, e-commerce, e-learning, not system specific, online banking, other

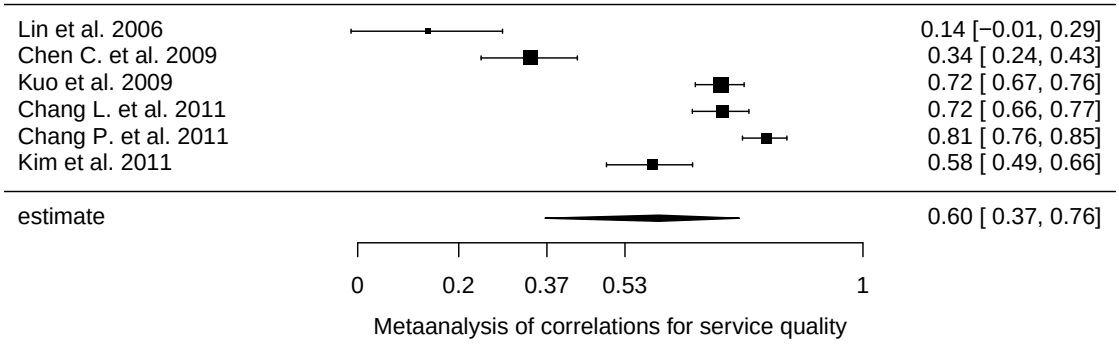
A.61 Service quality

ASMA category system appraisal

definition The quality of the service provided by a system to the user (our definition)

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.160$

heterogeneity measures $I^2 = 97.29\%$, $\tau^2 = 0.13$



relationship to satisfaction measured in 14 publications, 6 times as a correlation coefficient [30, 32, 31, 119, 135, 139, 173, 182, 189, 205, 211, 210, 212, 108]

measured for following types of system business information system, e-commerce, e-government, mobile system, not system specific, online banking, online community, other, telecommunication network, unspecified website

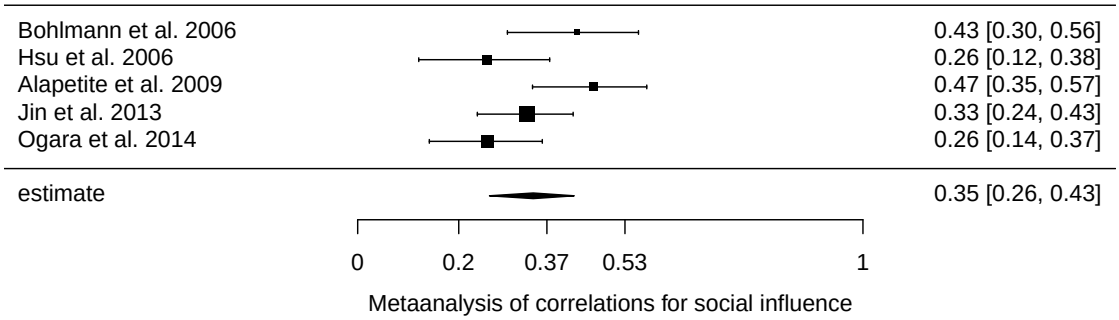
A.62 Social influence

ASMA category context mutable properties

definition Users' social beliefs about product usage are created by their interactions with other users in an organization or group. (from a primary study [102])

effect strength for satisfaction relationship Medium. Difference from lower strength is significant, $p = 0.009$

heterogeneity measures $I^2 = 61.5\%$, $\tau^2 = 0.01$



relationship to satisfaction measured in 6 publications, 5 times as a correlation coefficient [6, 22, 92, 102, 151, 199]

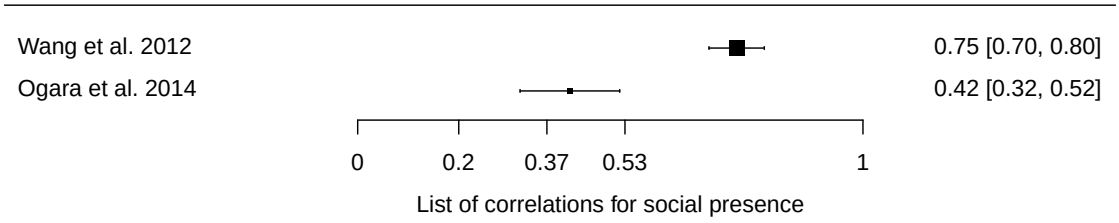
measured for following types of system e-commerce, e-learning, mobile system, not system specific, online banking, other, telecommunication network

A.63 Social presence

ASMA category system stable properties

definition the degree to which a communication channel facilitates awareness of communications partners and interpersonal relationship during interaction (from a primary study [151])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 2 times as a correlation coefficient [151, 188]

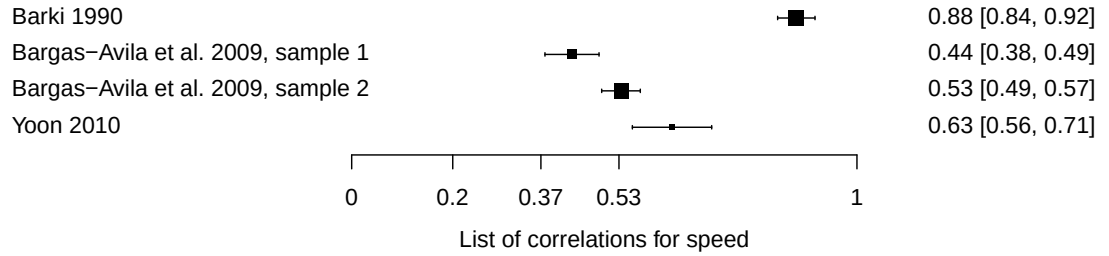
measured for following types of system telecommunication network

A.64 Speed

ASMA category system activity

definition The rate at which tasks can be done with the system (based on a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 4 times as a correlation coefficient [12, 14, 201]

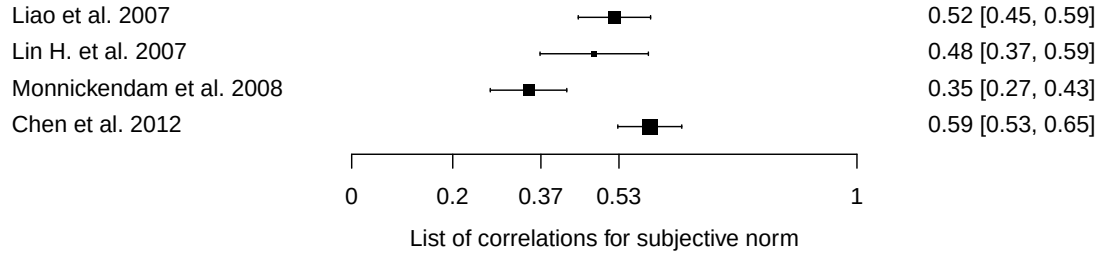
measured for following types of system business information system, online banking, other

A.65 Subjective norm

ASMA category context stable properties

definition the perceived social pressure to perform a particular behavior. (from a primary study [34])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 4 times as a correlation coefficient [34, 130, 134, 148]

measured for following types of system e-commerce, e-learning, not system specific, online community

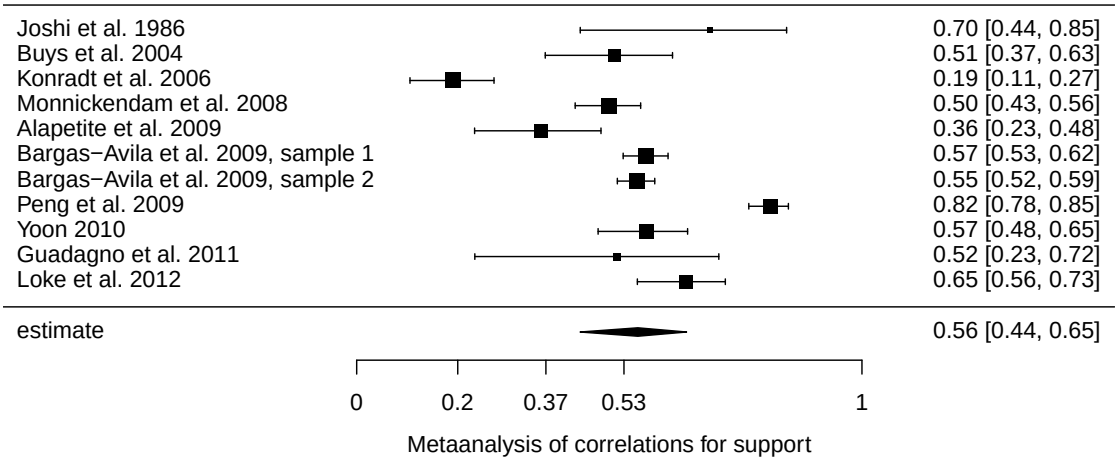
A.66 Support

ASMA category context stable properties

definition The amount of help provided to users when they need it (our definition)

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.021$

heterogeneity measures $I^2 = 95.42\%$, $\tau^2 = 0.06$



relationship to satisfaction measured in 10 publications, 11 times as a correlation coefficient [6, 12, 25, 75, 104, 141, 148, 155, 201, 116]

measured for following types of system business information system, e-learning, entertainment, not system specific, online banking, other

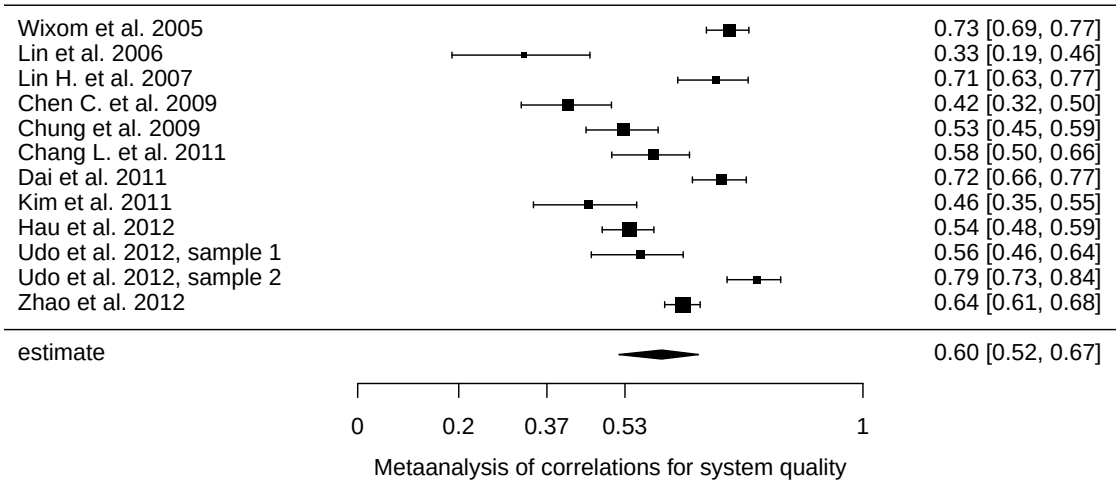
A.67 System quality

ASMA category system appraisal

definition System quality measures the desired characteristics of a system (from a primary study [135])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p < 0.001$

heterogeneity measures $I^2 = 94.18\%$, $\tau^2 = 0.04$



relationship to satisfaction measured in 22 publications, 12 times as a correlation coefficient [32, 31, 42, 48, 105, 117, 135, 134, 173, 181, 182, 189, 191, 197, 205, 211, 212, 108, 195, 84, 206, 126]

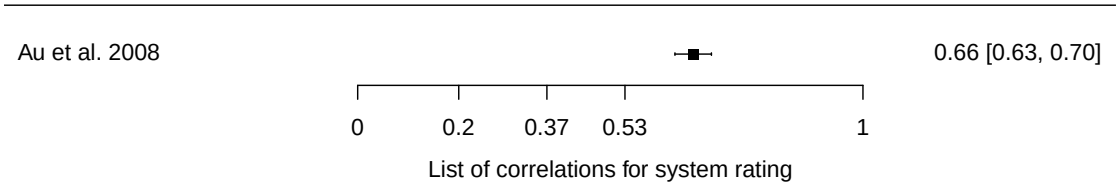
measured for following types of system blog, business information system, e-commerce, e-government, e-learning, mobile system, not system specific, online banking, online community, other, unspecified website

A.68 System rating

ASMA category system appraisal

definition A user's overall impression of the system expressed on an ordinal scale (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 1 times as a correlation coefficient [5, 10, 21]

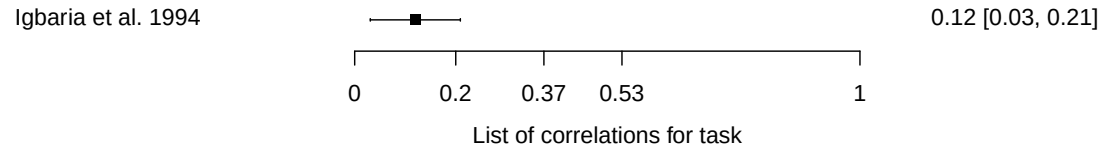
measured for following types of system business information system, e-learning

A.69 Task

ASMA category context mutable properties

definition a piece of work that the user needs to complete (based on a dictionary definition [147])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 2 publications, 1 times as a correlation coefficient [13, 96]

measured for following types of system e-government, e-learning, not system specific

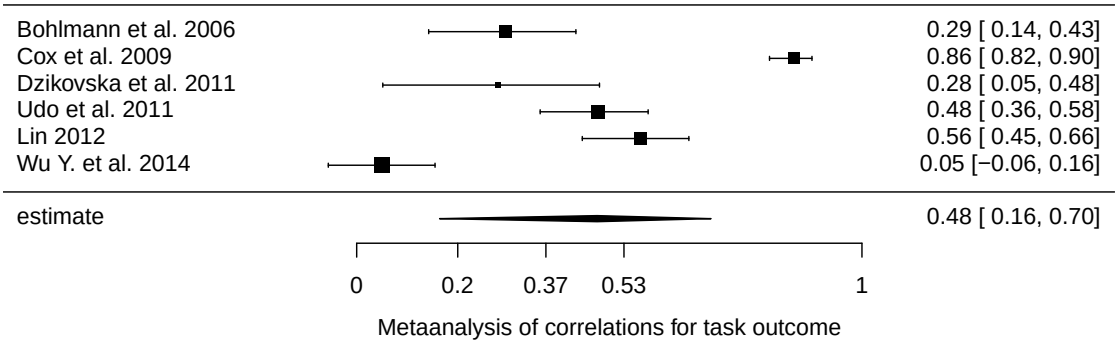
A.70 Task outcome

ASMA category context mutable properties

definition Denotes whether the user was able to complete the task successfully. (our definition)

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.245$

heterogeneity measures $I^2 = 96.93\%$, $\tau^2 = 0.19$



relationship to satisfaction measured in 6 publications, 6 times as a correlation coefficient

[22, 47, 57, 138, 180, 197]

measured for following types of system e-learning, online community, other

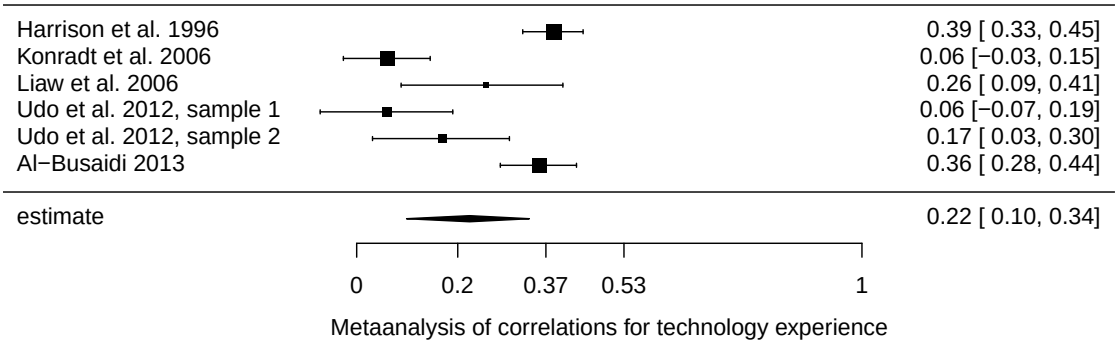
A.71 Technology experience

ASMA category user stable properties

definition The individual's exposure to the technology and the skills and abilities that s/he gains through using a technology (from a primary study [4])

effect strength for satisfaction relationship Weak. Difference from lower strength is significant, $p = 0.005$

heterogeneity measures $I^2 = 88.96\%$, $\tau^2 = 0.02$



relationship to satisfaction measured in 4 publications, 6 times as a correlation coefficient [4, 79, 131, 116]

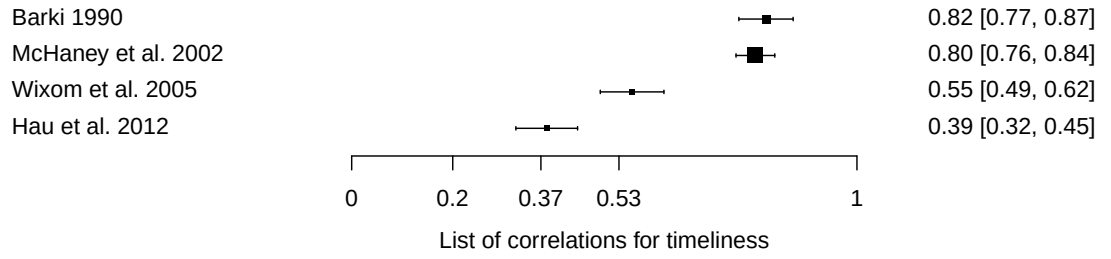
measured for following types of system business information system, e-government, e-learning, not system specific, other

A.72 Timeliness

ASMA category information mutable properties

definition The ability of a system to provide information quickly when it is needed (our definition)

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 4 times as a correlation coefficient [14, 146, 195, 84]

measured for following types of system business information system, mobile system, not system specific

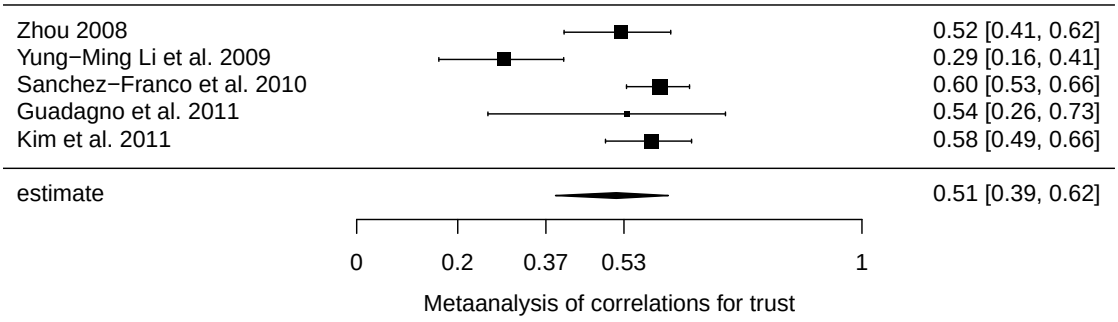
A.73 Trust

ASMA category user mutable properties

definition the willingness to rely on the system, having confidence in the system (from a primary study [202])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.111$

heterogeneity measures $I^2 = 81.8\%$, $\tau^2 = 0.02$



relationship to satisfaction measured in 8 publications, 5 times as a correlation coefficient [75, 173, 202, 207, 209, 108, 162, 126]

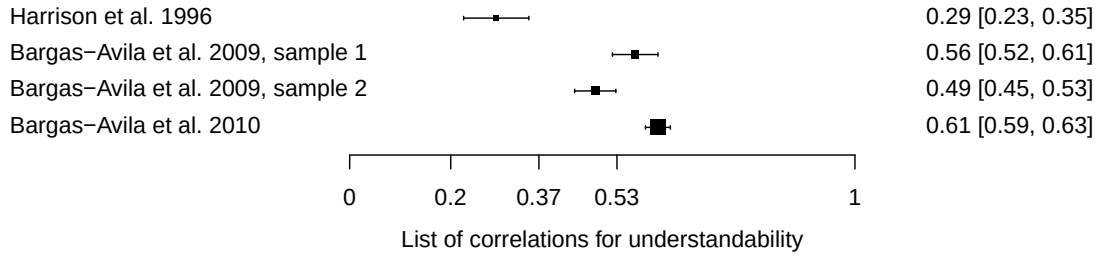
measured for following types of system e-commerce, entertainment, mobile system, online banking, other, unspecified website

A.74 Understandability

ASMA category information stable properties

definition Concerned with such issues as clearness and goodness of the information. (from a primary study [37])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 3 publications, 4 times as a correlation coefficient [12, 13, 79]

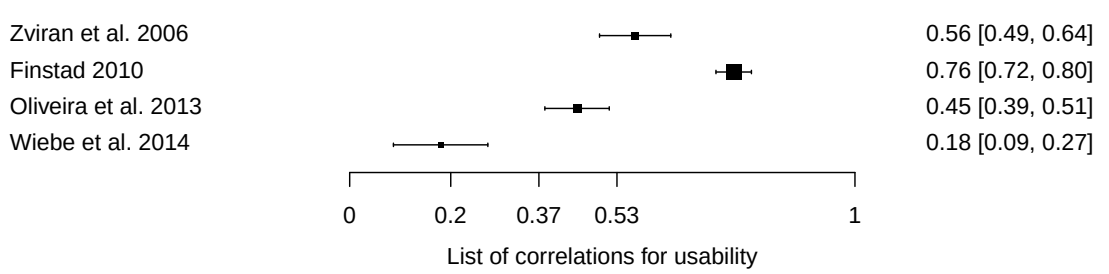
measured for following types of system blog, business information system, e-government, not system specific

A.75 Usability

ASMA category system appraisal

definition the ability of the user to use the thing to carry out a task successfully (from a specialized source [179])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 6 publications, 4 times as a correlation coefficient [27, 65, 152, 193, 213, 127]

measured for following types of system blog, e-commerce, entertainment, mobile system, not system specific

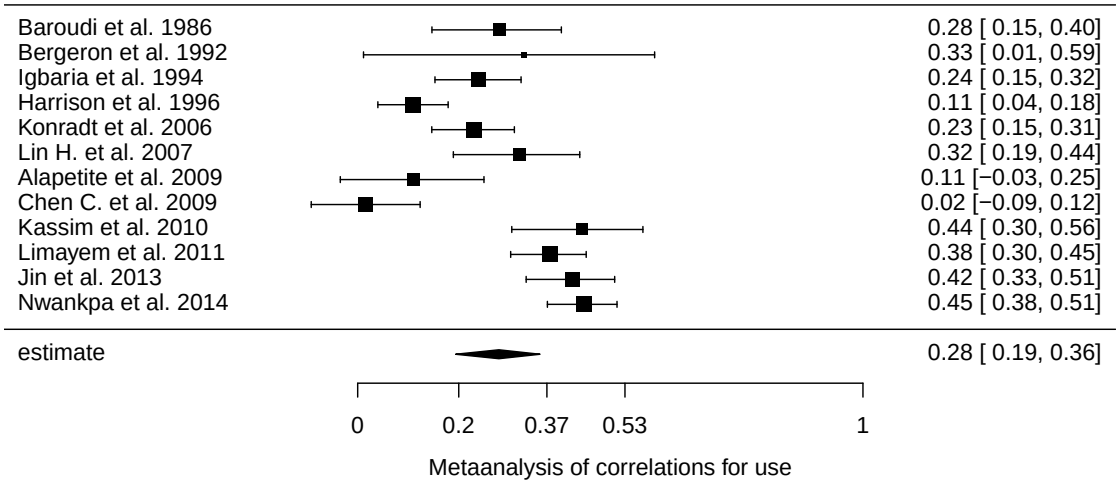
A.76 Use

ASMA category user activity

definition The frequency or total amount of time spent interacting with the system (our definition)

effect strength for satisfaction relationship Weak. Difference from lower strength is significant, $p < 0.001$

heterogeneity measures $I^2 = 87.84\%$, $\tau^2 = 0.02$



relationship to satisfaction measured in 17 publications, 12 times as a correlation coefficient [6, 8, 15, 18, 21, 32, 79, 86, 96, 102, 107, 133, 134, 150, 173, 182, 116]

measured for following types of system business information system, e-commerce, e-government, e-learning, mobile system, not system specific, online banking, online community, other

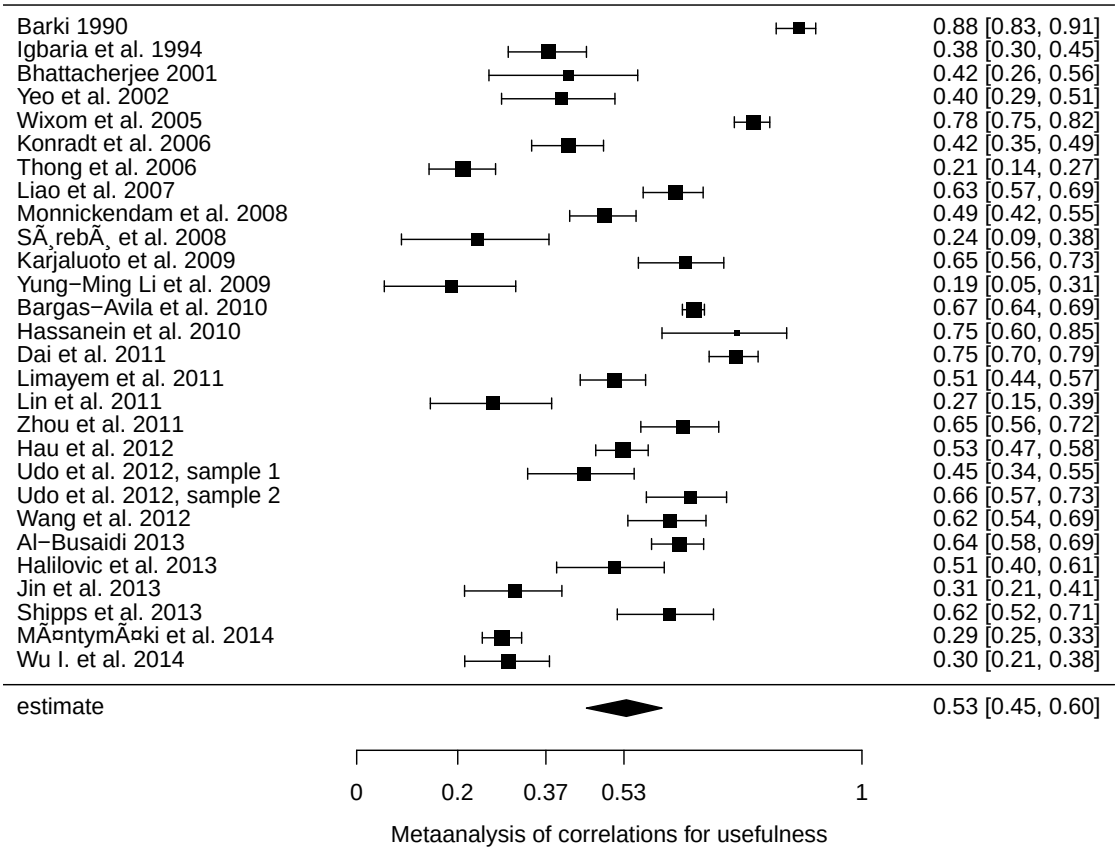
A.77 Usefulness

ASMA category system appraisal

definition 'the user's subjective probability that using a specific application system will increase his or her performance (based on a primary study [77])

effect strength for satisfaction relationship Strong. Difference from lower strength is significant, $p = 0.002$

heterogeneity measures $I^2 = 97.13\%$, $\tau^2 = 0.08$



relationship to satisfaction measured in 37 publications, 28 times as a correlation coefficient [13, 14, 20, 26, 48, 74, 77, 82, 86, 96, 102, 105, 130, 133, 137, 139, 148, 154, 171, 172, 176, 181, 200, 202, 209, 208, 212, 144, 132, 195, 188, 169, 35, 196, 106, 84, 116]

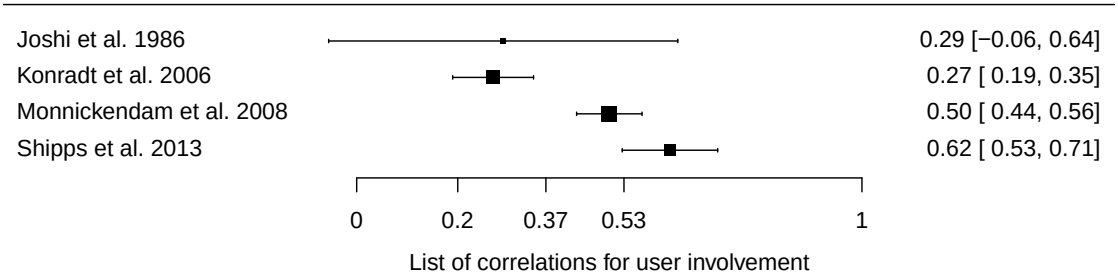
measured for following types of system blog, business information system, e-commerce, e-government, e-learning, entertainment, mobile services, mobile system, not system specific, online banking, online community, other, telecommunication network, unspecified website

A.78 User involvement

ASMA category context stable properties

definition psychological state of the individual, defined as the importance and personal relevance of a system to a use (from a specialized source [1])

effect strength for satisfaction relationship No metaanalysis conducted



relationship to satisfaction measured in 4 publications, 4 times as a correlation coefficient [104, 148, 169, 116]

measured for following types of system business information system, not system specific, online community

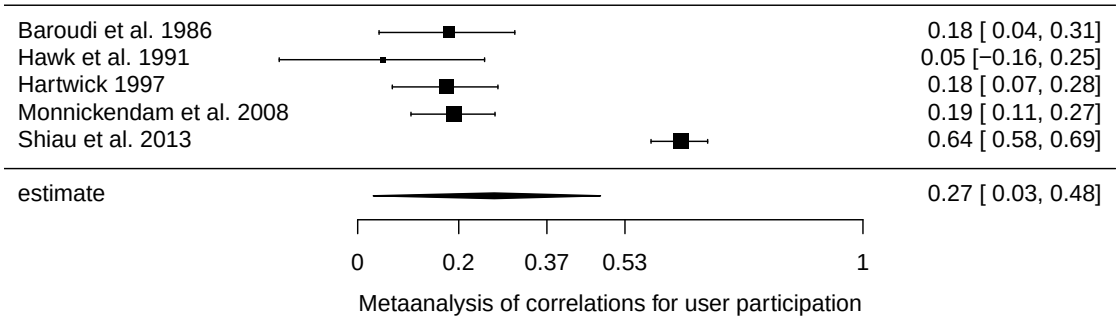
A.79 User participation

ASMA category context stable properties

definition behaviors and activities users perform in the system development Proces (from a specialized source [1])

effect strength for satisfaction relationship Weak. Difference from lower strength is not significant, $p = 0.111$

heterogeneity measures $I^2 = 95.61\%$, $\tau^2 = 0.07$



relationship to satisfaction measured in 6 publications, 5 times as a correlation coefficient [15, 81, 85, 148, 132, 168]

measured for following types of system blog, business information system, not system specific, online community

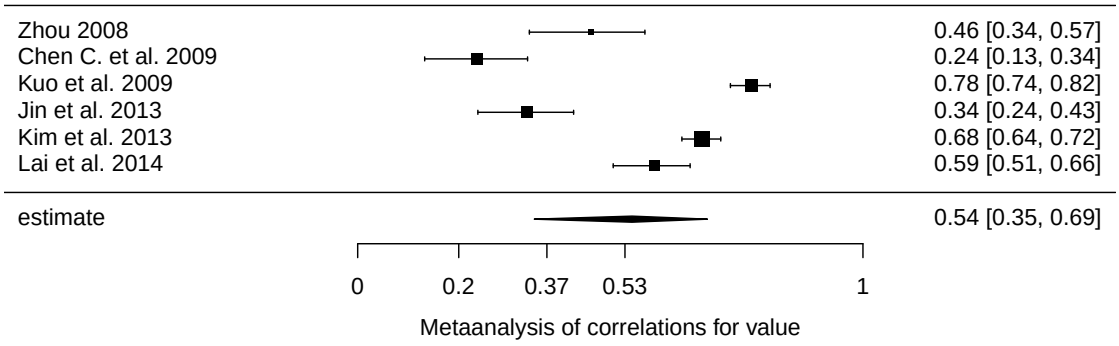
A.80 Value

ASMA category context mutable properties

definition a combined assessment of the perceived sacrifice in obtaining the service (e.g., price) and the benefits received (from a primary study [22])

effect strength for satisfaction relationship Strong. Difference from lower strength is not significant, $p = 0.168$

heterogeneity measures $I^2 = 96.99\%$, $\tau^2 = 0.09$



relationship to satisfaction measured in 9 publications, 6 times as a correlation coefficient [22, 32, 102, 119, 128, 187, 207, 120, 109]

measured for following types of system business information system, e-commerce, e-learning, mobile system, other

B Questionnaires used in our studies

In this appendix, we reproduce the questionnaires we used in our empirical studies. They are given in the following order:

- Live experiment at RefsQ 2012, used as an exploratory study to construct a model of anticipated satisfaction. The reproduced questionnaire shows the user task requirements format.
- Student experiment at the University of Heidelberg, used to validate the model of anticipated satisfaction. The reproduced questionnaire shows the user task requirements format.
- Casino study at the DKFZ, used to test the MUSA method's validity.
- Student experiment at the DKFZ, used to test the MUSA method's usability.

We reproduce the questionnaires exactly as given to the participants. The student experiment at University of Heidelberg and the Casino study at the DKFZ were conducted in German language, and the questionnaire reproduction here is not translated.

Live Experiment RefsQ 2012
Questionnaire

Part I
General questions

This part of the questionnaire includes some general questions about your experience in requirements engineering. Please answer the questions below. The experiment is anonymous; you are not required to provide a name or other identifying information.

Question 1
I have done the following work in requirements engineering (please cross all appropriate boxes):

- ☐ Research as a graduate student (Master, Ph.D.)
- ☐ Research at an academic institution (e.g. professor or junior professor)
- ☐ Research at a private company
- ☐ Teaching requirements engineering
- ☐ RE work at a software vendor
- ☐ RE work at a software customer
- ☐ RE work as a consultant
- ☐ Other (please specify):

Question 2 I have _____ years of experience working in requirements engineering.

Question 3 I rate my experience level in requirements engineering as:

beginner		intermediate		expert
☐	☐	☐	☐	☐

Question 4 In my practice in requirements engineering, I have used following format(s) to document requirements:

- ☐ User tasks
- ☐ User stories
- ☐ Sentence templates (”The system shall...”)
- ☐ Requirements without any format (plain text, evtl. ad-hoc structured)
- ☐ Other (please specify):

Question 5 In my practice in requirements engineering, the main format I use for documenting requirements is:

Part II
Requirements document

This part of the questionnaire contains requirements for the software product Receipt Manager. Imagine that you plan to buy the product when it is released. The software vendor knows you are interested in the finalized product and asks you for feedback as a key user.
The vendor provided some requirements, which are printed below. Please read them carefully and try to imagine the resulting software in detail. Then answer the questions about what you expect of the finished software based on the requirements.

Task 1	Digitize receipt
Start:	User has received one or more receipts
End:	The data from the receipts are archived
Subtasks	**Solution**
Import receipt	1 The user can import a picture of a receipt. 2 The user can prepare a picture for easier recognition. 3 The user can let the system recognize the text in an imported receipt picture. 4 The user doesnt predefine a tag, the system guesses a suitable tag and applies it to each expense item on a recognized receipt
Check and correct receipt data	5 The user can change any part of the recognized content. 6 The user can change the tag of each expense item separately.
Archive data	7 alternative 1: The user doesnt have to save anything. The data is kept on the system vendors server (cloud storage) and each change is automatically saved as it is made. 8 alternative 2: The user can save the receipt data locally. The user has to trigger the saving process. 9 The user can export the receipt data.
Task 2	View expenses report
Start:	User has received one or more receipts
End:	The data from the receipts are archived
Subtasks	**Solution**
Select the input for the report	10 The user can select the receipts to be used in a report. 11 The user can filter the receipt list or search it, for finding relevant receipts. 12 The user can choose from templates for different types of report. 13 The user inputs parameter needed for the report, depending on the type of report. For example, he/she inputs a month for a report which shows expenses for a given month.
Produce report	14 The user can read the report onscreen, it is shown in a printer-friendly layout. It contains data calculated by the system for the specified report. 15 The user can print the report 16 The user can export the report data.

Feature 1 The user can import a picture of a receipt.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ unsure ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ slightly important ☐ ☐ not important

Feature 2 The user can prepare a picture for easier recognition.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ unsure ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ slightly important ☐ ☐ not important

Feature 3 The user can let the system recognize the text in an imported receipt picture.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ unsure ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ slightly important ☐ ☐ not important

Feature 4 The user doesnt predefine a tag, the system guesses a suitable tag and applies it to each expense item on a recognized receipt

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ unsure ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ slightly important ☐ ☐ not important

Feature 5 The user can change any part of the recognized content.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ unsure ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ slightly important ☐ ☐ not important

Feature 6 The user can change the tag of each expense item separately.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ unsure ☐ not at all

2. I think this feature is ... for the way I will work with the system.
   very important ☐ ☐ ☐ ☐ ☐ not important

Feature 7 alternative 1: The user doesnt have to save anything. The data is kept on the system vendors server (cloud storage) and each change is automatically saved as it is made.

1. I can envision a way this feature will be implemented.
   really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system.
   very important ☐ ☐ ☐ ☐ ☐ not important

Feature 8 alternative 2: The user can save the receipt data locally. The user has to trigger the saving process.

1. I can envision a way this feature will be implemented.
   really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system.
   very important ☐ ☐ ☐ ☐ ☐ not important

Feature 9 The user can export the receipt data.

1. I can envision a way this feature will be implemented.
   really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system.
   very important ☐ ☐ ☐ ☐ ☐ not important

Feature 10 The user can select the receipts to be used in a report.

1. I can envision a way this feature will be implemented.
   really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system.
   very important ☐ ☐ ☐ ☐ ☐ not important

Feature 11 The user can filter the receipt list or search it, for finding relevant receipts.

1. I can envision a way this feature will be implemented.
   really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system.
   very important ☐ ☐ ☐ ☐ ☐ not important

Feature 12 The user can choose from templates for different types of report.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ ☐ ☐ not important

Feature 13 The user inputs parameter needed for the report, depending on the type of report. For example, he/she inputs a month for a report which shows expenses for a given month.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ ☐ ☐ not important

Feature 14 The user can read the report onscreen, it is shown in a printer-friendly layout. It contains data calculated by the system for the specified report.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ ☐ ☐ not important

Feature 15 The user can print the report

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ ☐ ☐ not important

Feature 16 The user can export the report data.

1. I can envision a way this feature will be implemented. — really well ☐ ☐ ☐ ☐ ☐ not at all
2. I think this feature is ... for the way I will work with the system. — very important ☐ ☐ ☐ ☐ ☐ not important

The software vendor asks you to prioritize the existing features. You have a total of 15 budget points. Please assign them to the features you feel are the most important ones. Mark each point in the space to the left of the features.
You can spend up to three points per feature, and you have to spend all fifteen points.

Task 1	Digitize receipt	
Start: **End:**	User has received one or more receipts The data from the receipts are archived	
Subtasks	**Solution**	
Import receipt	1 The user can import a picture of a receipt. 2 The user can prepare a picture for easier recognition. 3 The user can let the system recognize the text in an imported receipt picture. 4 The user doesnt predefine a tag, the system guesses a suitable tag and applies it to each expense item on a recognized receipt.	
Check and correct receipt data	5 The user can change any part of the recognized content. 6 The user can change the tag of each expense item separately.	
Archive data	7 alternative 1: The user doesnt have to save anything. The data is kept on the system vendors server (cloud storage) and each change is automatically saved as it is made. 8 alternative 2: The user can save the receipt data locally. The user has to trigger the saving process. 9 The user can export the receipt data.	

Task 2	View expenses report	
Start: **End:**	User needs information on expenses User has viewed the report and, optionally, printed or exported it.	
Subtasks	**Solution**	
Select the input for the report	10 The user can select the receipts to be used in a report. 11 The user can filter the receipt list or search it, for finding relevant receipts. 12 The user can choose from templates for different types of report. 13 The user inputs parameter needed for the report, depending on the type of report. For example, he/she inputs a month for a report which shows expenses for a given month.	
Produce report	14 The user can read the report on-screen, it is shown in a printer-friendly layout. It contains data calculated by the system for the specified report. 15 The user can print the report 16 The user can export the report data.	

Part III
Assessment of implemented features

Now the software product ReceiptManager is finished. You watch a demonstration which shows you how the product works. Please provide feedback based on the implemented software you saw in the demonstration.
For your convenience, you are shown a few features at a time, then given some time to answer, then another demonstration, etc., until you have seen all features.

Demonstration 1

Feature 1 The user can import a picture of a receipt.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 2 The user can prepare a picture for easier recognition.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 3 The user can let the system recognize the text in an imported receipt picture.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 4 The user doesnt predefine a tag, the system guesses a suitable tag and applies it to each expense item on a recognized receipt

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Demonstration 2

Feature 5 The user can change any part of the recognized content.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 6 The user can change the tag of each expense item separately.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Demonstration 3

Feature 7 alternative 1: The user doesnt have to save anything. The data is kept on the system vendors server (cloud storage) and each change is automatically saved as it is made.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 8 alternative 2: The user can save the receipt data locally. The user has to trigger the saving process.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 9 The user can export the receipt data.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all

3. The implemented feature differs from what I had envisioned in following ways :

Demonstration 4

Feature 10 The user can select the receipts to be used in a report.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 11 The user can filter the receipt list or search it, for finding relevant receipts.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 12 The user can choose from templates for different types of report.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 13 The user inputs parameter needed for the report, depending on the type of report. For example, he/she inputs a month for a report which shows expenses for a given month.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Demonstration 6

Feature 14 The user can read the report onscreen, it is shown in a printer-friendly layout. It contains data calculated by the system for the specified report.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 15 The user can print the report

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Feature 16 The user can export the report data.

1. I like the feature the way it is implemented now — a lot ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
2. The feature implementation corresponds to what I had envisioned — very well ☐ ☐ ☐ ☐ ☐ somewhat ☐ not at all
3. The implemented feature differs from what I had envisioned in following ways :

Part IV
The End

Thank you for participating in this experiment. We will evaluate the data during the conference and present the results at the end of the conference. Use the space below to give us feedback on this survey and on the experiment as a whole. Include critique, comments, questions, or anything else you feel we should know.

Rumi Proynova, Prof. Barbara Paech & the RefsQ organizers

My feedback:

Studie Featureverständnis
Fragebogen

Teil I
Allgemeine Fragen

Dieser Teil des Fragebogens enthält einige allgemeine Fragen zu Ihrem Hintergrund und Vorerfahrungen. Bitte beantworten Sie diese Fragen möglichst exakt. Die Studie ist anonym; Sie müssen keinen Namen oder andere Informationen angeben, mit denen Sie identifiziert werden können.

Frage 1 Ich studiere das Fach _____ im _____ Semester.

Frage 2 Meine Erfahrungen mit dem Fach Informatik (Mehrfachauswahl zulässig):

Ich habe Erfahrungen aus einem Studiengang mit mindestens 50% Anteil an Informatikvorlesungen:

☐ ich habe darin einen Abschluss (Bachelor, Master oder Diplom)

☐ ich studiere das gerade oder habe es früher studiert

Ich habe aus der Schule oder einem Studiengang einzelne Erfahrungen mit Informatikvorlesungen (z.B. Programmierung, Datenbanken, nicht IT-Fertigkeiten zu Office-Produkten:)

☐ Schule

☐ Studiengang (auch Nebenfach). Bitte geben Sie den Namen an: _____

Frage 3 Meine Erfahrungen mit Softwareentwicklung (Mehrfachauswahl zulässig):

☐ ich habe Software entwickelt, die ich selbst benutzt habe.

Ich war auf der Entwicklerseite an der Entwicklung einer Software beteiligt, die andere benutzt haben. Dabei:

☐ habe ich Anforderungsdokumente geschrieben, die andere verwendet haben.

☐ war ich an Entwurf oder Programmierung beteiligt.

☐ war ich als Tester oder in einer anderen Rolle an der Entwicklung beteiligt.

Ich war auf der Kundenseite an der Entwicklung einer Software beteiligt, die andere benutzt haben. Dabei:

☐ habe ich Anforderungsdokumente geschrieben.

☐ habe ich die Software benutzt.

☐ war ich in einer anderen Rolle beteiligt.

Frage 4 Ich schätze meine Erfahrung im Umgang mit Anforderungen wie folgt ein:

sehr viel		etwas		keine
☐	☐	☐	☐	☐

Teil II
Anforderungsbeschreibung

Dieser Teil des Fragebogens enthält Anforderungen (Beschreibungen von Features) für die Software *Receipt Manager*. Stellen Sie sich vor, dass diese Software in Planung ist. Sie haben vor, die Software zu kaufen, sobald sie fertig ist. Die Softwareentwicklerinnen wissen, dass Sie an die Software interessiert sind, und möchten Ihr Feedback als zukünftiger Benutzer.
Die Softwareentwicklerinnen haben eine Liste von Features entworfen, die Sie hier lesen können. Bitte lesen Sie die Featurebeschreibung aufmerksam durch und versuchen Sie, sich die resultierende Software im Detail vorzustellen. Bitte fangen Sie mit den Fragen erst an, nachdem Sie die Featurebeschreibung so gut wie möglich verstanden haben.

Task 1	Quittung digitalisieren
Start:	Der Benutzer hat eine Quittung erhalten
End:	Die Daten aus der Quittung sind archiviert
Subtasks	**Lösung**
Quittung importieren	1 Das System importiert ein Bild der Quittung. 2 Das System macht es leicht, ein Bild für das Erkennen vorzubereiten. 3 Das System erkennt den Text im Bild. 4 Das System rät ein Schlüsselwort und markiert damit jeden Posten auf der Quittung. Der Benutzer muss keine Schlüsselwörter vordefinieren.
Quittungsdaten prüfen und korrigieren	5 Das System erlaubt dem Benutzer, alles im erkannten Text zu korrigieren. 6 Das System erlaubt dem Benutzer, jedes geratene Schlüsselwort separat zu ändern.
Daten archivieren	7a) Option 1: Das System archiviert die Daten auf den Servern der Betreiber (Cloud Speicherung). Lokales Speichern ist nicht möglich. Alle Daten werden automatisch gespeichert. 7b) Option 2: Anstelle von Cloud Speicherung, werden die Daten lokal gespeichert. Der Benutzer muss das Speichern anstoßen. 8 Das System bietet das Exportieren der Daten aus der Quittung an.

Task 2	Bericht der Ausgaben anschauen
Start:	Der Benutzer möchte sich über Ausgaben informieren
End:	Die relevante Information wurde angesehen und möglicherweise gedruckt.
Subtasks	**Lösung**
Eingabedaten für den Bericht auswählen	9 Das System ermöglicht die Auswahl der Quittungen, die für einen Bericht benutzt werden sollen, aus einer Liste. 10 Das System bietet mehrere Such- und Filteroptionen an, um relevante Quittungen in der Liste zu finden. 11 Das System bietet Schablonen für unterschiedliche Arten von Berichten an. 12 Das System erlaubt dem Benutzer, Parameter für den Bericht einzugeben, z. B. den Monat für einen Bericht, der die Ausgaben für einen bestimmten Monat zeigt.
Bericht erstellen	13 Das System verarbeitet die Daten, die für den Bericht benötigt werden, und zeigt am Bildschirm einen drucker-freundlich formatierten Bericht. 14 Das System erlaubt dem Benutzer, den Bericht auszudrucken. 15 Das System erlaubt dem Benutzer, die Daten aus dem Bericht zu exportieren.

Nachdem Sie die Features verstanden haben, beantworten Sie bitte die folgenden Fragen dazu.

Feature 1 Das System importiert ein Bild der Quittung.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 2 Das System macht es leicht, ein Bild für das Erkennen vorzubereiten.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 3 Das System erkennt den Text im Bild.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 4 Das System rät ein Schlüsselwort und markiert damit jeden Posten auf der Quittung. Der Benutzer muss keine Schlüsselwörter vordefinieren.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 5 Das System erlaubt dem Benutzer, alles im erkannten Text zu korrigieren.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 6 Das System erlaubt dem Benutzer, jedes geratene Schlüsselwort separat zu ändern.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 7(a) Option 1: Das System archiviert die Daten auf den Servern der Betreiber (Cloud Speicherung). Lokales Speichern ist nicht möglich. Alle Daten werden automatisch gespeichert.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 7(b) Option 2: Anstelle von Cloud Speicherung, werden die Daten lokal gespeichert. Der Benutzer muss das Speichern anstoßen.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 8 Das System bietet das Exportieren der Daten aus der Quittung an.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 9 Das System ermöglicht die Auswahl der Quittungen, die für einen Bericht benutzt werden sollen, aus einer Liste.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 10 Das System bietet mehrere Such- und Filteroptionen an, um relevante Quittungen in der Liste zu finden.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 11 Das System bietet Schablonen für unterschiedliche Arten von Berichten an.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 12 Das System erlaubt dem Benutzer, Parameter für den Bericht einzugeben, z. B. den Monat für einen Bericht, der die Ausgaben für einen bestimmten Monat zeigt.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 13 Das System verarbeitet die Daten, die für den Bericht benötigt werden, und zeigt am Bildschirm einen drucker-freundlich formatierten Bericht.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 14 Das System erlaubt dem Benutzer, den Bericht auszudrucken.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Feature 15 Das System erlaubt dem Benutzer, die Daten aus dem Bericht zu exportieren.

1. Ich kann mir ... vorstellen, wie dieses Feature aussehen kann in der fertigen Software.

klar		vage		gar nicht
☐	☐	☐	☐	☐

2. Wenn das Feature umgesetzt ist, werde ich die folgende Einschätzung haben:

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

3. Ich kann mir ... vorstellen, warum dieses Feature nötig ist und wofür ich es benutzen werde.

klar		vage		gar nicht
☐	☐	☐	☐	☐

Die Softwareentwicklerinnen bitten Sie, die Features zu priorisieren. Sie haben insgesamt 15 Budgetpunkte. Bitte vergeben Sie sie an die Features, die Sie für die wichtigsten halten. Vermerken Sie die Punkte in den Kästchen neben den Features. Sie können bis zu 3 Punkte pro Feature vergeben, und Sie müssen alle 15 Punkte verwenden.

Task 1	Quittung digitalisieren
Start:	Der Benutzer hat eine Quittung erhalten
End:	Die Daten aus der Quittung sind archiviert
Subtasks	**Lösung**
Quittung importieren	☐☐☐ 1 Das System importiert ein Bild der Quittung. ☐☐☐ 2 Das System macht es leicht, ein Bild für das Erkennen vorzubereiten. ☐☐☐ 3 Das System erkennt den Text im Bild. ☐☐☐ 4 Das System rät ein Schlüsselwort und markiert damit jeden Posten auf der Quittung. Der Benutzer muss keine Schlüsselwörter vordefinieren.
Quittungsdaten prüfen und korrigieren	☐☐☐ 5 Das System erlaubt dem Benutzer, alles im erkannten Text zu korrigieren. ☐☐☐ 6 Das System erlaubt dem Benutzer, jedes geratene Schlüsselwort separat zu ändern.
Daten archivieren	☐☐☐ 7a) Option 1: Das System archiviert die Daten auf den Servern der Betreiber (Cloud Speicherung). Lokales Speichern ist nicht möglich. Alle Daten werden automatisch gespeichert. ☐☐☐ 7b) Option 2: Anstelle von Cloud Speicherung, werden die Daten lokal gespeichert. Der Benutzer muss das Speichern anstoßen. ☐☐☐ 8 Das System bietet das Exportieren der Daten aus der Quittung an.

Task 2	Bericht der Ausgaben anschauen
Start:	Der Benutzer möchte sich über Ausgaben informieren
End:	Die relevante Information wurde angesehen und möglicherweise gedruckt.
Subtasks	**Lösung**
Eingabedaten für den Bericht auswählen	☐☐☐ 9 Das System ermöglicht die Auswahl der Quittungen, die für einen Bericht benutzt werden sollen, aus einer Liste. ☐☐☐ 10 Das System bietet mehrere Such- und Filteroptionen an, um relevante Quittungen in der Liste zu finden. ☐☐☐ 11 Das System bietet Schablonen für unterschiedliche Arten von Berichten an. ☐☐☐ 12 Das System erlaubt dem Benutzer, Parameter für den Bericht einzugeben, z. B. den Monat für einen Bericht, der die Ausgaben für einen bestimmten Monat zeigt.
Bericht erstellen	☐☐☐ 13 Das System verarbeitet die Daten, die für den Bericht benötigt werden, und zeigt am Bildschirm einen drucker-freundlich formatierten Bericht. ☐☐☐ 14 Das System erlaubt dem Benutzer, den Bericht auszudrucken. ☐☐☐ 15 Das System erlaubt dem Benutzer, die Daten aus dem Bericht zu exportieren.

Teil III

Bewertung der umgesetzten Features

Die Software *Receipt Manager* ist fertiggestellt worden. Die Softwareentwicklerinnen führen Ihnen die Software vor. Bitte beantworten Sie folgende Fragen zu den Features, so wie sie umgesetzt sind. Um das Beantworten zu erleichtern, werden Ihnen zunächst einige Features gezeigt, dann haben Sie Zeit, dazu die Antworten anzugeben. Dann geht es weiter mit den nächsten 2-3 Features, etc.

Vorführung 1

Feature 1 Das System importiert ein Bild der Quittung.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 2 Das System macht es leicht, ein Bild für das Erkennen vorzubereiten.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 3 Das System erkennt den Text im Bild.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 4 Das System rät ein Schlüsselwort und markiert damit jeden Posten auf der Quittung. Der Benutzer muss keine Schlüsselwörter vordefinieren.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Vorführung 2

Feature 5 Das System erlaubt dem Benutzer, alles im erkannten Text zu korrigieren.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 6 Das System erlaubt dem Benutzer, jedes geratene Schlüsselwort separat zu ändern.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Vorführung 3

Feature 7 a) Option 1: Das System archiviert die Daten auf den Servern der Betreiber (Cloud Speicherung). Lokales Speichern ist nicht möglich. Alle Daten werden automatisch gespeichert.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 7 b) Option 2: Anstelle von Cloud Speicherung, werden die Daten lokal gespeichert. Der Benutzer muss das Speichern anstoßen.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 8 Das System bietet das Exportieren der Daten aus der Quittung an.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Vorführung 4

Feature 9 Das System ermöglicht die Auswahl der Quittungen, die für einen Bericht benutzt werden sollen, aus einer Liste.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 10 Das System bietet mehrere Such- und Filteroptionen an, um relevante Quittungen in der Liste zu finden.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 11 Das System bietet Schablonen für unterschiedliche Arten von Berichten an.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einiger-maßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 12 Das System erlaubt dem Benutzer, Parameter für den Bericht einzugeben, z. B. den Monat für einen Bericht, der die Ausgaben für einen bestimmten Monat zeigt.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Vorführung 5

Feature 13 Das System verarbeitet die Daten, die für den Bericht benötigt werden, und zeigt am Bildschirm einen drucker-freundlich formatierten Bericht.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 14 Das System erlaubt dem Benutzer, den Bericht auszudrucken.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Feature 15 Das System erlaubt dem Benutzer, die Daten aus dem Bericht zu exportieren.

1. Die Umsetzung dieses Features entspricht ... meiner früheren Vorstellung.

sehr gut		einigermaßen		gar nicht
☐	☐	☐	☐	☐

2. Das umgesetzte Feature unterscheidet sich von meiner Vorstellung in folgender Art und Weise:

3. Meine Einschätzung für das Feature, so wie es jetzt umgesetzt ist, ist.

ich mag es		ich bin indifferent		ich mag es nicht
☐	☐	☐	☐	☐

4. Ich habe schon ein ähnliches Feature in einer anderen Software benutzt.

ja, sehr ähnlich		ja, entfernt ähnlich		nein
☐	☐	☐	☐	☐

Teil IV
Abschlussfragen

Frage 5 Haben Sie Ideen, wie man die Featurebeschreibung verbessern kann, so dass Sie sich die Features leichter vorstellen können?

Frage 6 Ich fand die Idee, diese Software zu benutzen:

großartig		langweilig		unsinnig
☐	☐	☐	☐	☐

Wir danken Ihnen für die Teilnahme an unserer Studie. Wir werden die Ergebnisse auswerten und für unsere Forschung verwenden. Wenn Sie Interesse haben, mehr zu den Ergebnissen zu erfahren, schicken Sie bitte eine E-Mail an Frau Proynova und Sie werden benachrichtigt, wenn die Ergebnisse publiziert werden.

Wenn Sie Feedback für uns haben - Anregungen, Anmerkungen, Eindrücke, oder was auch immer Sie uns sagen wollen - schreiben Sie es bitte auf dieses Blatt, bzw. verwenden Sie die Rückseite.
Rumyana Proynova, Barbara Paech & das VaReMed Team

Mein Feedback:

Mit der Abgabe dieses Fragebogens erklären Sie sich einverstanden, dass Ihre Antworten für Forschungszwecke gespeichert und ausgewertet werden. Ihre Anonymität ist dabei gewährleistet. Die im Rahmen dieser Studie angefallenen persönlichen Daten (z. B. Ihre E-Mail Addresse) werden ausschließlich für die Organisation dieser Studie verwendet, nicht an Dritte weitergegeben, und nicht mit Ihren Antworten in Verbindung gebracht.

Dieser Fragebogen ist Teil des Projekts VaReMed (Value Based Requirements for Medical Software), das von der DFG finanziert wird.

C Report from a MUSA application

We validated our method by applying it in two studies, the Casino study and the MITO study. There was no report for the MITO study, since the stakeholders decided that they do not need one. For the Casino study, we made measurements of both actual and anticipated satisfaction to allow comparison of the two measurements, and the original report reflected that. This made it a poor illustration for a normal MUSA application. Thus we produced a report taking into account only the MUSA measurement of anticipated satisfaction. This can be used as an example of how we intend such reports to be produced.

Anticipated satisfaction with the Casino Catering application

Rumyana Proynova
Deutsches Krebsforschungszentrum

August 21, 2017

Abstract

1 Study

This study concerns the software system for catering orders that was developed at the DKFZ in 2016. With this system, DKFZ employees can order food from the Casino for events they are organizing. The system was developed in-house and is available to all employees. It replaces an old paper-based process that was earlier used for catering orders.

1.1 Goals of the study

The study measures the users' anticipated satisfaction with the system. It has following research questions:

1. Is there a need to provide a "special orders" button, or is it superfluous?
2. How do users feel about the controversial feature of e-mail updates on an order status?

With the results of this study, the DKFZ management gets an insight into the quality of the internally developed software. For the Casino, it is an opportunity to understand the desires and attitudes of their customers and to react to them. The developers can get a better understanding of how the system is used and how the users relate to it.

1.2 Conducting the study

The study consisted of a questionnaire that measures anticipated satisfaction. The questionnaire was made available online to employees, using DKFZ's Limesurvey server. A pseudonym was used to conceal the participants' identity.

The questionnaire was made available online while the system was still in development. It was announced in the internal e-mail list, which invited employees to participate in the study. This means that all participants were potential users, even though some of them were not acquainted with the catering ordering process.

The questionnaire contained demographic data, questions about the central data record (the order form) and two representative features of the system, the "special requests" features and the "email updates" feature. It concluded with a section with questions about the system as a whole, and gave participants the opportunity to share their opinion in free text.

2 Results

2.1 Participants

There were 39 participants who filled the questionnaire. They were free to skip questions, but there was not much missing data, and there was no evidence for systematic non-answering.

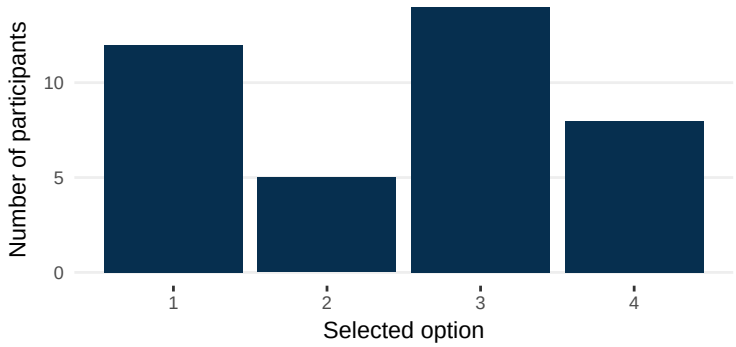


Figure 1: Participants by frequency of placing catering orders. 1 - never placed an order, 2 - less than 1x per year, 3 - 1 to 10 orders per year, 4 - over 10 orders per year.

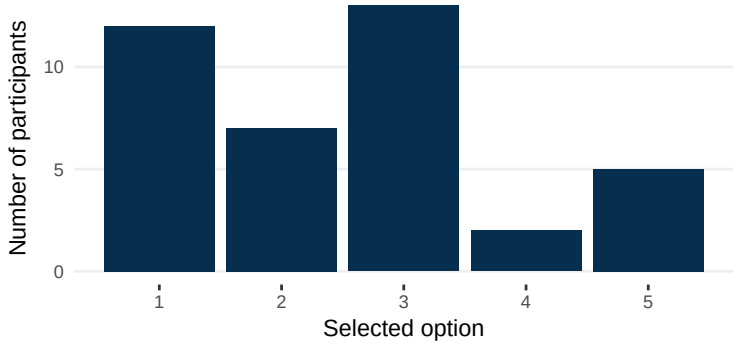


Figure 2: Participants by frequency of placing catering orders. 1 - never placed an order, 2 - less than 1x per year, 3 - 1 to 10 orders per year, 4 - over 10 orders per year.

Most participants place catering orders moderately often (1 to 10 times per year), and there was also a sizable group of 12 participants who have not placed orders before. A bar graph in figure 1 shows the distribution of participants depending on their order frequency.

We also asked the participants if they have used a similar software system previously. In this case, the answers were unanimous, and nobody had had experience with a system of this type.

The third demographic question was to ask participants whether they enjoy experimenting with technology. Most of them chose either the highest or the neutral option, with few answers in the middle or on the low part of the scale. The distribution is plotted in figure 2.

2.2 Research question 1

The first research question is about the "special orders" functionality. There were three variables measured for it - *usefulness*, *satisfaction* and *perceived understanding*. A comparison between the functionality and the values of these variables overall is shown in 3.

The functionality's usefulness is good, the same as overall usefulness. Satisfaction is excellent in both the functionality and overall score. The only difference is in perceived understanding, where the functionality is only good, whilve overall it is excellent. The functionality is not polarizing in any of the variables.

We can conclude that the functionality has a good standing with the users, and does not substantially differ from the overall evaluation. There are no reasons for change or removal from the specification. As the perceived understanding is somewhat lower than overall, it may be warranted to try making it more intuitive for users, e.g. by exploring the effect of different wordings for the button.

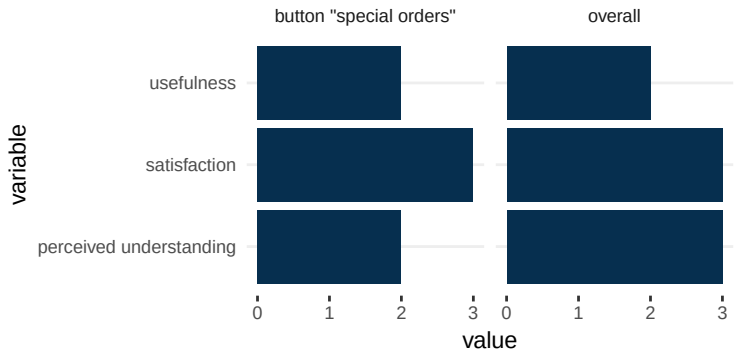


Figure 3: Anticipated satisfaction with the "special orders" functionality

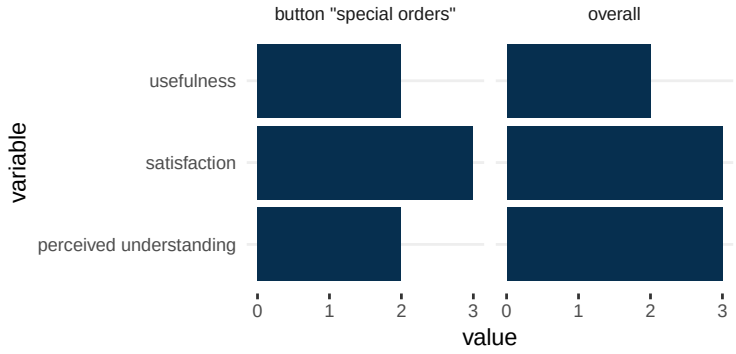


Figure 4: Anticipated satisfaction with the "updates by email" functionality

2.3 Research question 2

The second research question asks about the controversiality of the functionality that sends status updates by emails.

The scores for that functionality and the overall scores are plotted in figure 4. The functionality scores excellent in all four categories. There is no evidence for controversy. Therefore the feature should be kept in the application as-is.

2.4 Further findings

Almost all variables in the study fell in the expected positive categories, that is, either *good* or *excellent*. An exception was *complexity*, which fell in the *bad* category. It is unclear whether this impression will hold after some use of the application, as the low score may be a side effect of being faced with something unfamiliar. There were no polarizing features, or unclear ones.

There were six free-text comments left (excluding one complaint of a technical issue with the survey software), and most of them contained suggestions for a new feature. These are given in table 1.

The suggestions vary from easy fixes (e.g. change the ordering of fields in the form) through new features which will require a change in process (e.g. letting the Casino decide how much food to deliver) to those that concerned features out of scope of the survey, e.g. billing. We suggest that the team goes through the list of suggested changes/new features and considers which ones can be included in the specification.

The overall satisfaction predicted by the survey is excellent. A summary graph of all answers is provided in figure 5.

I would change the ordering of the data in the order form, because the date and time belong to the event and not to the order! So better this ordering: Division - Event - Date - Time - Location - Number of participants - Phone number - Second contact person with phone number - Desired time for pickup. Explanation: In our division, the order is mostly made through the office manager, but the pickup is sometimes made by the colleague who organizes the meeting. This would be the second contact person and it would be good, if he is informed about the order status. It would also be advantageous if one could enter the desired pickup time in the order heading.
For the function email updates, the user should decide if he wants to use the function and he should be able to deie, if he only wants updates for changes in the order (e.g. amounts, items) or also changes in order status
hello, I looked at the survey: 1. There is no way to enter time and location for the pickup 2. What is a "functionality"? 3. Because this wasn't clear to me and the questions were irrelevant, I stopped answering.
I believe, the trick is to find the right balance between what one can pick and what is offered, for example the amount of extra china. It is important for me to be able to make a planned/actual correction. I order 10 bottles of water, only 6 get drunk, how does this get documented and billed. I didn't understand this from the survey.
Is the order followable? This wasn't clear. It should be possible to not only give a number of the desired article, but also a checkbox for the desired item as such and the amount for e.g. pots of coffee should be left to the Casino.
Our order contains usually the point "service personnel" and the amount "as needed". Are these components available in the new program under the special orders? Should one not list a contact person in the order form? This would be advantageous for our internal processing, since multiple colleagues make orders for different events.

Table 1: Free-text feedback

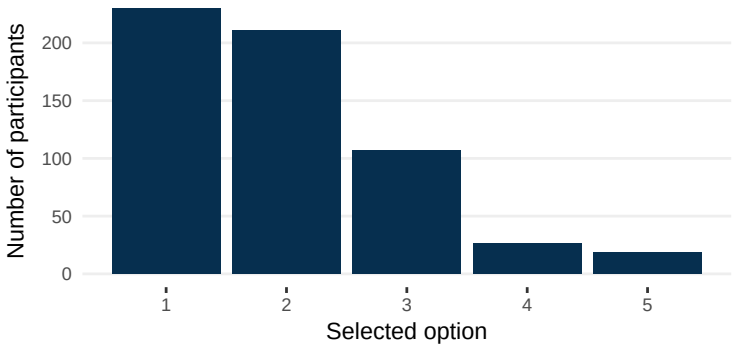


Figure 5: Distribution of all answers combined

2.5 Conclusions

In this survey, 39 participants were asked for their anticipated satisfaction with the new Casino catering application. The overall satisfaction was excellent, and there were no unusual effects found. The two features analyzed in detail did not present any special issues. Some new requirements were suggested in free text.

Our recommendation is to keep the feature "special orders", but consider a new wording for the button. The feature "email updates" can be optionally improved by allowing the users to set the desired email behavior as a preference, as suggested by one user in the freetext, but that requirement should be validated with other sources, since the majority of users are very happy with the feature as it is. The team should go through all freetext answers and elicit potential new requirements from them.

3 Acknowledgements

The study was planned and conducted by Rumyana Proynova (employee IT Core Facility), who also wrote the report. Claudia Galuschka (Leader Databases at the IT core facility) helped in the planning phase and supported the conduction. Martin Hauschild (Leader Casino) made the study possible and supported the questionnaire design and the implementation. Nicolas Helfrich (employee IT Core facility) implemented the application and was involved in the survey planning. We had additional support from Ms Malinowski (Casino), Olga Daum (Data protection office), Bernd Raseman and Monika Gai (Support LimeSurvey software) and the Personalrat, with whose approval we conducted the study.