Human-Centered Crowd Feedback for Information Systems Development

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Saskia Haug, M.Sc.

Tag der mündlichen Prüfung:	29.01.2024
Referent:	Prof. Dr. Alexander Mädche
Korreferent:	Prof. Dr. Petra Nieken
Karlsruhe	Januar 2024

Acknowledgments

First and foremost, I want to thank Prof. Alexander Mädche, my mentor, and PhD supervisor, for his unwavering guidance, inspiration, and valuable feedback throughout my studies. At the beginning of this journey, I was unsure of what to expect, but I quickly learned how fortunate I was to have Alexander as my supervisor. His open-door policy for my various questions, ideas, and concerns, coupled with his encouragement to submit my research to prestigious outlets, has been a cornerstone of my PhD experience. I am immensely grateful for his support. Likewise, I want to thank Prof. Petra Nieken, JProf. Julian Thimme, and Prof. Ingrid Ott for their willingness to serve on my PhD committee. I had a great time discussing my studies with you!

I want to express my gratitude to my colleagues at the Institute of Information Systems and Marketing (IISM), particularly those in the Human-Centered Systems Lab. Thank you for assisting with pre-tests, providing critical feedback during our research meetings, and sharing countless coffee and tea breaks that offered much-needed respite and laughter!

I am extremely grateful and delighted to have always had the support of my friends! am grateful for the bonds that have remained strong from the early days of my university study, and I am grateful to those I met during my later studies. Thank you for the good times, your support, and your constant presence.

Finally, my heartfelt thanks also go to my parents and my family. This dissertation is dedicated to my mother, Claudia Haug, and my father, Alfred Haug. Your unconditional love, endless support, and unwavering belief in my abilities have been the foundation for everything I've achieved. Thank you for walking this path with me, every step of the way.

Saskia Haug

Karlsruhe, Germany

January 2024

Abstract

User involvement in information system (IS) development is crucial for project success and user satisfaction. Traditional evaluation methods like usability tests and focus groups are often limited by scale, cost, and time constraints. Crowdsourcing offers a scalable, cost-effective, and timely alternative for gathering user feedback during IS development. Despite its potential, the feedback provider's perspective is often overlooked, raising concerns about meeting their needs and optimizing feedback quality. This dissertation explores crowd-feedback systems through a human-centered design process to balance the needs of feedback providers with the requester's goal of high-quality feedback collection. The dissertation tackles two main challenges: the need to better understand design feature impacts in crowd-feedback systems and the failure of existing crowd-feedback systems to collect user feedback in a real context. The research includes five studies delivering four innovative design solutions, evaluating their impacts, and understanding different stakeholder perceptions. The first study maps out the crowd feedback landscape, identifying key inputs, design features, crowdsourcing configurations, and effects. Based on this, it proposes multiple future research directions. Building upon these foundations, the second study designs 'Feeasy', a human-centered crowd-feedback system, and evaluates the impact of its features on feedback providers and feedback outcomes. The third study develops a configuration system enabling the customization of feedback requests. Addressing the second challenge, the fourth study tests integrating user and crowd feedback through the 'CrowdSurfer,' a browser extension, noting that feedback tasks embedded in regular internet use are seen as less effort but may reduce feedback quality. The fifth study introduces preference-based personalization in microtasking, assessing its influence on crowdworker performance and offering design guidelines for personalized microtasks. This research contributes to the fields of human-computer interaction and information systems by detailing crowd-feedback system design and investigating personalized crowdsourcing systems in real

contexts. The findings provide actionable design knowledge for practitioners to improve IS evaluations' scalability and human-centricity. The four developed artifacts apply this knowledge to real-world crowd feedback scenarios, aiming to improve IS evaluations by focusing on stakeholders' needs.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike's Information Criterium
ANOVA	Analysis of Variance
ART	Aligned Rank Transform
CEFR	Common European Framework of Reference for Languages
DR	Design Rationale
DSR	Design Science Research
EUD	End-User Development
GfeW	German Association for Experimental Economic Research
HCD	Human-Centered Design
HCI	Human-Computer Interaction
IS	Information Systems
LLM	Large-Language Model
MANOVA	Multivariate Analysis of Variance
MPI	Multitasking Preference Inventory
MTurk	Amazon Mechanical Turk
RQ	Research Question
SIMS	Situational Intrinsic Motivation Scale
SLR	Systematic Literature Review
SWOT	Strengths-Weaknesses-Opportunities-Threats
TIME	Theory of Interactive Media Effects
UCD	User-Centered Design

UI ..... User Interface

UX ..... User Experience

1. Introduction [1]

1.1. Motivation

Involving users in the information systems (IS) development process is crucial and is known to have positive impacts on various outcome dimensions such as IS project success (Harris & Weistroffer, 2009), user satisfaction (McKeen & Guimaraes, 1997), and user acceptance (Ives & Olson, 1984). Neglecting users' involvement in the IS development process can lead to reduced user satisfaction and, ultimately, project failure (Hsu et al., 2013). The Human-Centered Design (HCD) approach offers a methodological framework for involving users and other stakeholders in the development process (ISO 9241-210, 2019; Vredenburg et al., 2002). The HCD process emphasizes the importance of involving users in key activities of IS design, including analysis, specification, design, and evaluation (Brhel et al., 2015; ISO 9241-210, 2019). One of the most challenging aspects of HCD is the continuous evaluation of possible design solutions with potential users (Brhel et al., 2015). According to the iterative approach of HCD, it is important to evaluate early sketches, mock-ups, and clickable prototypes of IS designs. This can be done in formative and summative evaluations. While formative evaluations are carried out during the process, summative evaluations happen at the end of the process to evaluate the outcomes (Scriven, 1991). Popular methods for evaluating IS designs are usability tests, interviews, and focus groups. Depending on the lifecycle phase of the IS and the evaluation goal, different methods are appropriate. For instance, in traditional lab-based usability testing, participants are required to perform tasks by interacting with the artifact. The goal is to measure the time and clicks users need to find the information and capture their thoughts to identify usability issues (Nielsen, 1994). Therefore, to conduct a usability test, at least some interactivity and information must be provided by the artifact. Due to the face-to-face character of usability tests, focus groups, and interviews, these techniques lack scalability. As they require human experts to guide or interview participants, they are also costly and time-consuming for designers and developers. At the same time, these methods also require potential users who are willing to participate in these studies, and companies often

[1] This chapter is based on the following studies: Haug and Maedche (2021a), Haug, Benke, and Maedche (2023), Haug, Benke, Fischer, and Maedche (2023), Haug, Sommerrock, et al. (2023), and Haug, Benke, Fischer, Walther, et al. (2023)

struggle to find a diverse group of potential users of the artifact (Mackay, 2004).

Another approach for IS evaluation is the collection of feedback, usually from existing users of the IS. This feedback can be requested by the designers and developers, in the following called feedback requesters, via feedback pop-ups that are integrated into the IS. Alternatively, feedback can also be provided by actual users when they share their opinions in online forums or app stores.

While these approaches in their traditional form are only applicable to IS that are already in use, crowdsourcing feedback is an emerging approach that is also applicable during the development of the IS through dedicated feedback studies. Crowdsourcing means the process of gathering information or input of a task or project from a large number of people, either paid or unpaid, typically via the internet (Howe, 2006). The benefits of crowdsourcing feedback are that it is scalable, relatively inexpensive, and can be applied during the development as well as usage of the IS. Therefore, it is very flexible which is also demonstrated by the various applications that exist in research e.g., chatbots (Choi et al., 2021), mobile apps (Ayalon & Toch, 2018, 2019), and reinforcement learning systems (de la Cruz et al., 2015). In crowd feedback, a large group of people who must not necessarily be actual or potential users of the system are asked for their opinion on an IS design or an actually running IS, like an online available website (Y. W. Wu & Bailey, 2016). Crowd feedback originally emerged from the graphical design domain, where it was used to replace and scale peer feedback (Wauck et al., 2017). Dedicated crowd-feedback systems were developed to collect feedback with a specific focus on graphics design (Luther, Pavel, et al., 2014; Xu & Bailey, 2014). In recent years more and more studies as well as crowd-feedback systems have been presented for the more general IS evaluation context (Luther, Tolentino, et al., 2015; Oppenlaender, Tiropanis, & Hosio, 2020). Various studies have been conducted to show the feasibility of crowdsourcing design feedback (e.g., Oppenlaender, Tiropanis, and Hosio, 2020; Y. W. Wu and Bailey, 2016; Yuan et al., 2016). These systems enable feedback requesters to provide users with structure and guidance for delivering targeted feedback. The resulting data from these systems comprise both quantitative and qualitative insights, including issues, praises, and ideas for further improvement from the user's perspective (Oppenlaender, Kuosmanen, et al., 2021). Consequently, crowd-feedback systems differ from established online forums as they collect structured feedback, which is more actionable compared to unstructured comments or aggregated individual preferences

found in online forums (Xu & Bailey, 2014). Existing research on crowd-feedback systems has demonstrated their ability to collect feedback of similar quality to that of design expert feedback (Y. W. Wu & Bailey, 2016; Yuan et al., 2016). Numerous studies have explored the positive impacts of specific design features in crowd-feedback systems on outcomes such as quality, scalability, and effort (Choi et al., 2021; Greenberg et al., 2015; Oppenlaender, Tiropanis, & Hosio, 2020). However, despite the promising potential of crowd-feedback systems to scale the IS evaluation process, existing crowd-feedback systems still face challenges that limit their applicability in practice. Although crowd feedback has already proven to be a scalable approach for IS evaluation, there are still challenges regarding the design of crowd-feedback systems and their application.

First, while existing studies show that crowdsourcing design feedback for different types of designs is feasible for overcoming issues of traditional evaluation methods and presenting diverse approaches, features, and designs, it does not explore their specific effects. This is limiting the prescriptive knowledge in this field of research. There is a gap between the systems that were developed in research and their application in practice. Existing systems in research were only developed for a specific use case and are not applicable to the evaluation of different designs in further scenarios. They apply quantitative or qualitative methods, including features like markers, categories, questionnaires, and direct manipulation, and evaluate many different types of systems like chatbots, websites, and static designs. Some of them are in early development stages, others already live. A general problem is that often design decisions are not made explicit. However, to give recommendations to practitioners and design crowd-feedback systems that can be applied to various use cases, prescriptive knowledge for designing crowd-feedback systems is necessary.

Second, the advantages of feedback pop-ups that are integrated into IS are that feedback providers are actual users and are in the context of use when providing feedback. Feedback pop-ups also allow a continuous collection of feedback that might show differences over time or after the release of new features. However, they often fail due to a lack of motivation from users to provide meaningful feedback. The feedback quality is often low and does not go beyond simple statements like ”I like it”. Existing crowd-feedback systems usually collect feedback only at a specific point in time (e.g., Y. W. Wu and Bailey, 2016), take users out of an actual context of use (e.g., Ayalon and Toch, 2019), or fail to provide good incentives beyond the possibility of an improved IS design or feeling involved in the

development process (e.g., Haukipuro et al., 2016). I argue that feedback systems that are able to bridge the gap between traditional feedback pop-ups and forums and existing crowd-feedback systems might be able to combine the best of both worlds and be able to collect inexpensive high-quality feedback from actual users at scale. One important aspect here is the motivation of crowdworkers to provide high-quality feedback. There is already much research exploring what crowdworkers value about tasks. Among others, they value autonomy, fairness, and transparency, but also having an impact and being proud of their work (Deng, Joshi, & Galliers, 2016). There is a recent rise in research on personalizing crowdwork to increase worker motivation and job performance, mainly according to cognitive styles (Paulino, Correia, Barroso, & Paredes, 2023). However, I argue that there is still a lack of research that investigates personalization according to further important characteristics of crowdworkers, such as polychronicity and social preferences as it is known that these characteristics also impact work-related behavior (Asghar, Gull, et al., 2020; Cassar, 2018).

In this thesis, I investigate how to improve the design of crowd-feedback systems following a human-centered design approach. The main goal of this thesis is to design systems that achieve the best results for feedback requesters and feedback providers by not only considering the feedback outcomes but also focusing on the perceptions of feedback providers. Specifically, I address the two design challenges discussed above: (i) to understand the effect of specific crowd-feedback system features to inform the design of future crowd-feedback systems and enable non-experts to apply crowd feedback for their unique use cases, and (ii) to make crowd feedback more real and bring actual users to provide high-quality design feedback. For the first challenge, I applied methodologies and insights from the human-computer interaction (HCI) discipline. For the second challenge, I combine approaches from HCI with the Design Science Research (DSR) paradigm to deliver a human-centered but also theory-driven design. With my studies, I contribute prescriptive knowledge in the form of design principles, rationales, recommendations, and goals. I also contribute descriptive knowledge in the form of a deeper understanding of the effects of different crowd-feedback system features and preference-based personalization in the context of casual microtasking. In the next paragraph, I describe the research gaps in more detail and derive research questions (RQs) that guided the studies of my thesis.

1.2. Research Gaps and Research Questions

This thesis explores human-centered crowd-feedback systems for IS design evaluation. I have already argued why evaluating IS early and continuously matters. I have explained what methods can be used and why traditional methods do not scale well. I have introduced crowd-feedback systems as an approach to collect design feedback in a scalable way. More specifically, I have argued that the design of crowd-feedback systems is currently not theory-grounded and needs to be better understood. Further, I have explained that a combination of crowd feedback and traditional user feedback could help to further enhance the resulting feedback. Also, I have argued, that personalizing crowdsourcing systems can have positive effects on job performance. Therefore, I formulate the following research question (RQ) to guide the research in my dissertation project:

Main Research Question: *How can human-centered crowd-feedback systems be designed to obtain high-quality feedback outcomes?*

To further decompose this overarching RQ, I will introduce multiple sub-RQs. The first goal is to understand what the current research on crowd-feedback systems encompasses. Crowd feedback is a comparably young field of research that evolved in the last decade and has its roots in the HCI research area. Previous research of Morschheuser et al. (2017), Pedersen et al. (2013), and Zuchowski et al. (2016) has provided conceptualizations of crowdsourcing systems in general, highlighting their components. Also Leicht (2018) provided a comprehensive overview of research on crowd testing. Although the transition from crowd testing to crowd feedback is fluid, to the best of my knowledge, before this thesis, there has not been a comprehensive overview of research on crowd-feedback systems published. Usually, crowd-feedback systems are tailored to specific use cases. Practitioners and researchers who want to build crowd-feedback systems need to know which features are available and how they impact the perceptions of users and the resulting feedback. To choose design features for crowd-feedback systems as educated decisions, one needs to connect these design features with context and desirable outcomes. Just recently, Alpar and Osterbrink (2018) highlighted the importance of including antecedents and outcomes in analyses of past research on crowdwork. As research now has identified multiple antecedents, design features, and outcomes that vary between all existing crowd-feedback systems, there is a need for a conceptual aggregation of crowd-feedback systems. Hence, I formulate my first sub-RQ as follows:

Research Question 1a: *How to conceptualize crowd feedback for IS development?*

Following a conceptualization, it is important to understand where current research has focused on to identify research gaps and develop new interesting research avenues. To do so, existing research must be gathered and analyzed using the conceptual framework that will be the result of sub-RQ1. Therefore, I formulate my second sub-RQ:

Research Question 1b: *What is the state-of-the-art of crowd feedback in IS development and what are future research directions?*

I investigate RQ1a and RQ1b by following a three-step approach. First, I conducted a systematic literature review (SLR) based on Webster and Watson (2002) and Kitchenham and Charters (2007), followed by a conceptualization of the found papers, and a cluster analysis. In the SLR, I identified 40 relevant papers in the context of crowd-feedback systems. The conceptualization reveals multiple research gaps. Among others, there is a lack of research on the effects of different feedback features on the resulting feedback, especially the perceptions of feedback providers (e.g., crowdworkers) are in existing studies often not considered. This also leads to a gap between research on crowd-feedback systems and their application in practice. The systems in the identified studies were usually built for a very specific use case or artifact and are therefore not applicable to other use cases (e.g., in practice). The cluster analysis led to three research streams on crowd-feedback systems.

After understanding the state of the art, my remaining studies built upon the identified research gaps and future research directions that I also outlined in my motivation. Namely, I introduced two main challenges that will guide my following four studies. The first challenge addresses the lack of prescriptive knowledge and the resulting limited applicability of crowd feedback. Without knowledge about the effects of individual design features of crowd-feedback systems, it is difficult to design effective crowd-feedback systems. Further, the majority of existing studies on crowd feedback neglected the perspective of users of the system. I argue, that understanding how crowdworkers perceive and interact with crowd-feedback systems is crucial for designing human-centered crowd-feedback systems that lead to high-quality feedback outcomes. Only the user engagement of feedback providers, e.g., crowdworkers, has been considered in a few studies (Hosseini et al., 2016; Robb, Padilla, Kalkreuter, & Chantler, 2015b; Robb, Padilla, Methven, et al., 2017; Snijders et al., 2015).

Before being able to analyze the effects of design features, I had to build an interactive crowd-feedback system following a human-centered design approach. Most of the existing crowd-feedback systems focus on static designs. In this study, we also wanted to address this research gap by designing a crowd-feedback system that can handle the specific requirements of evaluating interactive designs like clickable prototypes. Therefore, I propose the following sub-RQs:

Research Question 2a: *How to design a human-centered crowd-feedback system to evaluate interactive designs?*

Research Question 2b: *How do different crowd-feedback system design features affect crowdworkers' perceptions and the resulting feedback quality and quantity?*

To investigate RQ2a and RQ2b, I designed and developed *Feeasy* an interactive crowd-feedback system. *Feeasy* is based on the results of a design study with ten participants who tested two different crowd-feedback artifacts and shared their experiences. *Feeasy* combines five feedback features that were requested by participants of the design study. These five features and a combination of them were consecutively evaluated in an online experiment with 210 participants. The results show that the features *scenarios* and *categories* are perceived as the most important features and that combining too many features overwhelms crowdworkers as it decreases the perceived ease of use.

Although research has demonstrated that crowd feedback is able to solve persistent problems of traditional evaluation methods, there is still a lack of application of it in practice. I assume this is due to a lack of skills of designers to design and build individualized crowd-feedback systems. The results regarding RQ2b include recommendations on how to use and combine different feedback features in crowd-feedback systems. Based on these results, it is possible to offer other researchers and practitioners guidance in applying crowd feedback. My goal is to make crowd-feedback systems not only easy to use but also easy to configure (Lieberman et al., 2006). Therefore, I seek to answer the following research question:

Research Question 3: *How to design a configuration system to support designers in creating effective customized crowd-feedback requests?*

I address RQ3 by developing a configuration system based on *Feeasy*. Based on 14 expert interviews and a literature review, I designed and developed the configuration system that

can be used to adapt an instantiation of *Feeasy* to specific use cases. In a focus group workshop (N = 10) I investigated how the system is perceived by experts by conducting a SWOT analysis. The results demonstrate that the approach is appreciated by experts. However, a good balance between flexibility and complexity needs to be found.

The second main challenge, that I am addressing with this thesis is the combination of crowd feedback and traditional user feedback. While crowd feedback offers the benefits of scalability, diversity, and reduced effort, it also comes with some issues. Feedback providers are not necessarily potential users of the system and are also not in a real usage scenario when providing feedback. To tackle these issues, it is necessary to take the crowdworkers back into an actual context of use when providing feedback, also called *in situ feedback* (Froehlich et al., 2007; Seyff, Ollmann, & Bortenschlager, 2014). This can be done by asking them for feedback when they are surfing on these websites anyway and thereby integrating the feedback tasks into their everyday internet surfing. The term *casual microtasking* describes the concept of the integration of microtasks into other primary tasks (Hahn et al., 2019). While integrating tasks into other primary activities has the potential to reduce the invisible work of crowdworkers' such as searching for tasks and switching between websites that are needed for the task, it could also negatively influence the work-life balance of crowdworkers. Based on these thoughts, I propose the following fourth research question.

Research Question 4: *How to design a system to collect in situ crowd feedback in the form of casual microtasking to improve the working conditions of crowdworkers and feedback quality?*

I address RQ4 by designing and developing the *CrowdSurfer*, a Google Chrome extension that allows the integration of feedback tasks into crowdworkers' daily internet surfing. The *CrowdSurfer* allows the feedback collection from actual users, who are at the same time crowdworkers, by showing feedback tasks in the form of pop-ups whenever a participating crowdworker enters a website that is requesting feedback. The design was developed via interviews with crowdworkers from which I derived three design rationales. I instantiated them in the *CrowdSurfer* and showed in an online experiment, that this approach is feasible, although the resulting feedback is significantly worse ($p < 0.01$) than in a traditional feedback survey. Still, crowdworkers perceived the payment with the *CrowdSurfer* as significantly fairer ($p < 0.05$) and perceived that they needed significantly less time to solve

the tasks ($p < 0.05$). I further provide qualitative insights through additional interviews with participants ($N = 12$).

One of the key insights of the interviews was that there seem to be two types of crowdworkers. Some like the integration of microtasks into other primary activities, while others prefer to be able to switch between a work and a private mode. Personalizing crowdsourcing systems is an emerging research stream to improve crowdsourcing outcomes. Currently, the main focus lies on cognitive personalization (Paulino, Correia, Barroso, & Paredes, 2023; Paulino, Correia, Guimarães, et al., 2022). However, there are many more characteristics of crowdworkers that can be used for personalization. According to the person-environment fit theory (Caplan, 1987; Edwards, Caplan, & Van Harrison, 1998), workers are more satisfied and perform better when the job environment addresses their needs, abilities, and preferences. In the case of casual microtasking, especially polychronicity, the preference for multitasking, and social preferences should be considered. For example, the initial design of the *CrowdSurfer* required multitasking and therefore was designed contrary to the preferences of monotaskers. As every crowdworker differs in their preferences, I argue that there is a need for preference-based personalization of casual microtasking systems like the *CrowdSurfer* to increase job performance. Therefore, I propose the following two research questions:

Research Question 5a: *How to design a preference-based personalized casual microtasking system to increase job performance?*

Research Question 5b: *How do preference-based personalizations in casual microtasking systems affect job performance?*

I address RQ5a by adapting the *CrowdSurfer* design to create the *MyCrowdSurfer* system. The *MyCrowdSurfer* is designed to collect accessibility feedback, more specifically alt-tags for images on Wikipedia. The *MyCrowdSurfer* offers personalized designs according to crowdworkers' preferences. The theory-driven design is based on the P-E Fit theory (Caplan, 1987; Edwards, Caplan, & Van Harrison, 1998) and the results of the previous study. To answer RQ5b, I conducted a longitudinal field study on Prolific. In the experimental study, I applied the two design instantiations that were derived in the first step to analyze the impact of preference-based personalization on crowdworkers' job performance. The results demonstrate that personalization according to crowdworkers' polychronicity could

slightly increase the quality of task outcomes, but has no impact on the number of submitted tasks. Further, personalization according to crowdworkers' altruism as one type of social preference leads to a significant decrease in job performance, including the submitted quantity and quality of tasks. Especially altruistic crowdworkers perform much worse when they are using a system that highlights the altruistic goal of the task. I assume this is caused by a complex relationship between crowdworkers' intrinsic and extrinsic motivation, explained by phenomena like the *overjustification effect* and *tainted altruism*. With the study, I contribute both prescriptive and descriptive knowledge to the body of research on personalized crowdworking.

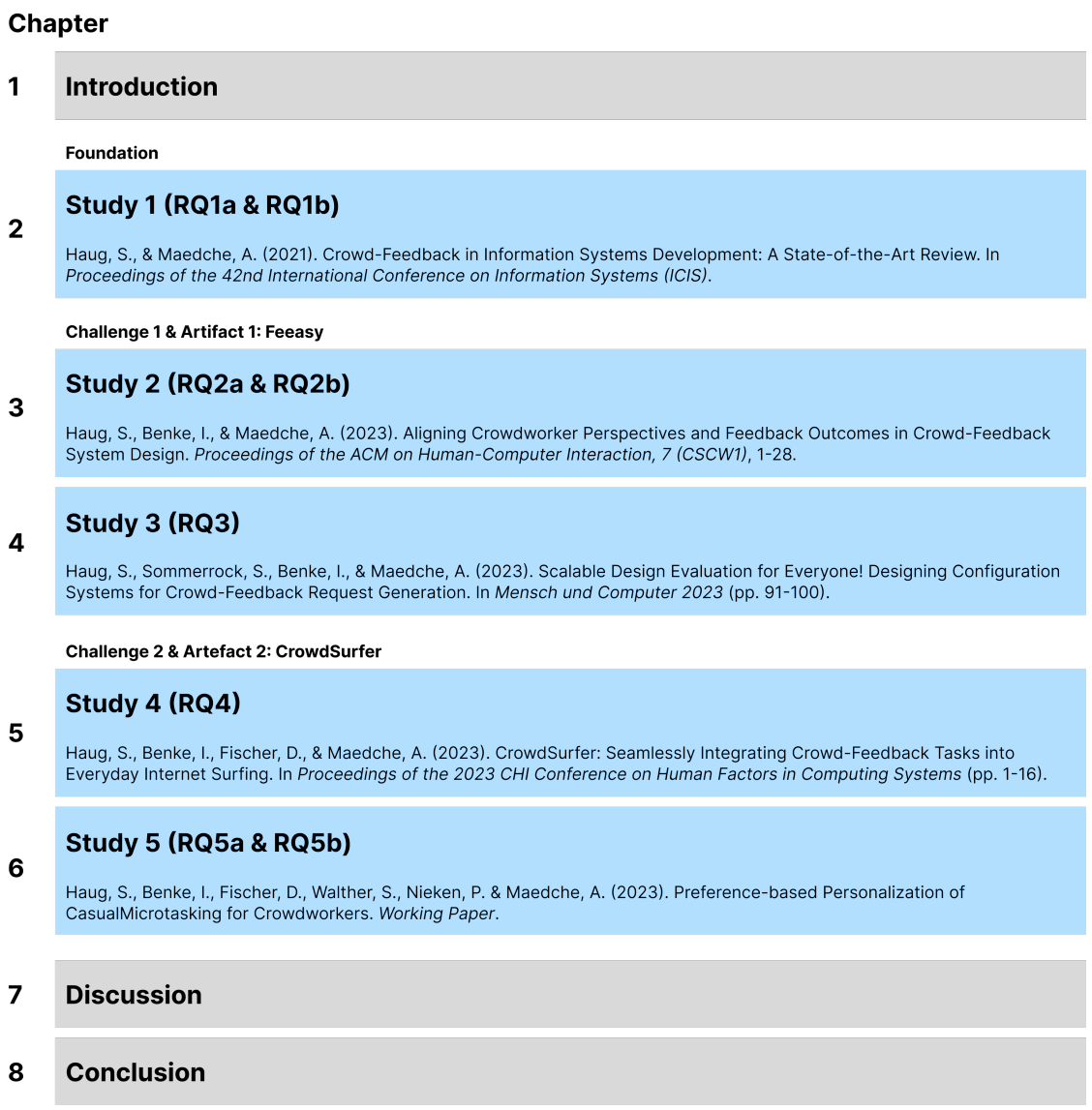


Figure 1.1.: Structure of the Thesis

1.3. Thesis Structure

In Figure 1.1 the structure of my cumulative thesis is illustrated. Chapter 1 motivates my thesis project, presents the research gaps and design challenges, describes the derived research questions, as well as explains the structure of this thesis. The research questions are addressed in five studies. These are described in **Chapters 2, 3, 4, 5,** and **6**. **Chapter 2** presents an SLR to understand the state-of-the-art crowd-feedback systems and to identify future research directions. **Chapter 3** presents *Feeasy* and investigates the effects of crowd-feedback design features on the perceptions of users and the resulting feedback quality and quantity. **Chapter 4** describes the design of a configuration system for crowd-feedback systems to enable non-experts to build and adapt crowd-feedback systems themselves. **Chapter 5** investigates the integration of crowd-feedback tasks into daily internet surfing of crowdworkers via the *CrowdSurfer*. **Chapter 6** investigates with the *MyCrowdSurfer* system the design of preference-personalized casual microtasking systems. **Chapter 7** summarizes the overall findings of this thesis and discusses practical and theoretical contributions. Finally, **Chapter 8** concludes this thesis.

2. Study I: A State-of-the-Art Review of Crowd-Feedback in Information Systems Development

2.1. Introduction

User involvement is of critical importance in the development of any information system (IS). It has a positive impact on IS project success (Harris & Weistroffer, 2009), user satisfaction (McKeen & Guimaraes, 1997), and IS acceptance (Ives & Olson, 1984). As such, not involving users in IS development may culminate in project failure (Hsu et al., 2013). User-centered design (UCD) is a prominent paradigm that provides methodological guidance for user involvement (Vredenburg et al., 2002). In the iterative UCD process, it is emphasized that users should be involved in all major activities of IS development, namely analysis, specification, design, and evaluation (Brhel et al., 2015; ISO 9241-210, 2019). Continuous evaluation of design solutions with potential users is a challenging activity in UCD (Brhel et al., 2015). In particular, designers have to cope with the scalability issues of traditional face-to-face methods when they aim to involve a diverse set of users in the development process. Thus, in recent years, crowdsourcing in user-centered IS evaluation has received growing interest in research (Alyahya, 2020; Leicht, 2018; K. Mao et al., 2017; Sarı et al., 2019) and practice (e.g., uTest, UsabilityHub).

Thereby, two main research streams can be identified: crowd testing and crowd feedback. While in crowd testing the crowd interacts with the IS in order to identify errors (Leicht, 2018), in crowd feedback, the crowd is asked for explicit, mostly verbal feedback, including opinions and perceptions of the IS design (Xu & Bailey, 2014). Interaction with the IS is thereby often not necessary and a textual description or screenshot of the user interface may be sufficient. However, the transition from crowd testing to crowd feedback is fluid, and there of course exist systems that include both approaches (e.g., Ayalon and Toch, 2019). While huge amounts of user feedback on existing systems are commonly provided in user forums and app stores (Pagano & Bruegge, 2013; Tizard et al., 2022; Yen, Dow, et al., 2016), collecting feedback during the IS development requires dedicated crowd-feedback systems and respective crowds. With these systems, feedback requesters can

provide users with structure and guidance for providing dedicated feedback. The outcomes of these systems then comprise quantitative and qualitative data, that contain valuable issues, praises, and ideas for further improvement from the user perspective (Oppenlaender, Kuosmanen, et al., 2021). Therefore, crowd-feedback systems differ from established online forums because the collected feedback is clearly structured and therefore more useful than unstructured comments or aggregated individual preferences collected in online forums (Xu & Bailey, 2014). Overall, existing research on crowd-feedback systems has shown to be able to collect feedback with a quality similar to expert feedback (Y. W. Wu & Bailey, 2016; Yuan et al., 2016). While existing approaches and systems on crowd-testing in IS were recently reviewed (Alyahya, 2020; Leicht, 2018), there is a lack of a systematic review of existing knowledge on crowd feedback and corresponding systems.

Although multiple studies already investigated the positive effects of design elements of crowd-feedback systems on outcomes like quality, scalability, and effort (Choi et al., 2021; Greenberg et al., 2015; Oppenlaender, Tiropanis, & Hosio, 2020), there is a lack of research that synthesizes the current body of knowledge. Specifically, the varying requirements of feedback collection endeavors via crowdsourcing are not well structured and conceptualized. This hinders researchers to identify possible directions for future research efforts in this emerging research stream. In this paper, we seek to focus on these challenges by addressing the following two research questions (RQ):

RQ1: *How to conceptualize crowd feedback for IS development?*

RQ2: *What is the state-of-the-art of crowd feedback in IS development and what are future research directions?*

In order to answer the RQs, we perform a systematic literature review (SLR) study based on Webster and Watson (2002) and Kitchenham and Charters (2007) that identifies and investigates 40 articles on crowd-feedback systems. Next, following the approach of Nickerson et al. (2013) and the grounded-theory method proposed by Wolfswinkel et al. (2013) we develop a conceptualization of crowd feedback in IS. By coding and analyzing the identified 40 articles, we provide a holistic overview of the state-of-the-art of crowd-feedback systems for IS evaluation. Based on a subsequent cluster analysis of the 40 articles, we identify three relevant research streams. Considering the previously provided state-of-the-art overview and the identified research streams, we present four avenues for future

research. We contribute to theory by providing a morphological box for crowd feedback for IS and a state-of-the-art review on this basis. This study additionally contributes to practice by supporting practitioners in applying crowd-feedback systems in IS development by outlining three research streams and related design characteristics. In the remainder of this paper, we first provide an overview of the theoretical foundations and illustrate the methods applied in this study. In section four the results of our study are presented. This is followed by a discussion and outline of future research directions in section five. Lastly, section six is the conclusion of our article.

2.2. Related Work and Conceptual Foundations

We first provide conceptual foundations on user-centered evaluation in IS development. Here our focus especially lies on the critical role of feedback in the user-centered evaluation and differentiating it from testing. This is followed by an introduction to crowdsourcing in IS development. Finally, we present existing conceptual frameworks on crowdsourcing systems.

2.2.1. User-Centered Evaluation in IS Development

There exist multiple methods to evaluate design solutions by involving the user, e.g., user interviews, focus groups, or usability testing (Gibbs, 1997; Vredenburg et al., 2002). A distinction is usually made between formative and summative evaluation, with formative evaluation being carried out during the process and summative evaluation at the end to evaluate the outcomes (Scriven, 1991). A general challenge is that formative and summative evaluation methods are time- and cost-intensive in their application and therefore lack scalability (Gibbs, 1997; Scholtz, 2001). This issue can be addressed by utilizing online crowds following a crowdsourcing paradigm.

The most fundamental method for evaluation is usability testing with potential users (Brhel et al., 2015; Nielsen, 1994). In usability testing, participants receive tasks that they must complete by interacting with the IS. During the test, the participant is observed and data, such as the time needed to complete a task and the number of required clicks, is measured. Additional subjective data is collected by recording participants' comments during the usage and optionally via a subsequent questionnaire (Nielsen, 1994, pp. 165-206). In comparison, in feedback collection, participants' comments, ratings, votes, and

markers are the only data that is collected. For existing systems, users provide feedback directly via pop-ups, app stores, or feedback forums (Almaliki et al., 2014). Feedback forums like Dribble are dedicated to providing feedback on designs and prototypes. The feedback quality on these platforms is often low and feedback requesters are not able to provide any guidance for users (Xu & Bailey, 2014). Therefore, dedicated feedback systems are required to collect valuable feedback that exceeds simplistic statements of "I like it" (Xu & Bailey, 2014). Many of these systems (e.g., Luther, Pavel, et al., 2014; Xu and Bailey, 2011) originated from the visual design domain. There, feedback, also known as design critique, is traditionally provided by peers to help designers understand how others perceive their designs (Yuan et al., 2016). Crowd-based feedback systems were initially developed to solve scalability issues of peer feedback (Wauck et al., 2017) and reach a more diverse crowd of feedback providers (Ma et al., 2015). The ongoing challenge, especially in the visual design domain, is to enable the non-expert crowd to provide feedback that is similar to feedback from design experts and investigate the differences between feedback from the crowd and peers (Wauck et al., 2017; Yuan et al., 2016).

2.2.2. Crowdsourcing in IS Development

Crowdsourcing is a method to outsource tasks to a large undefined crowd of people (Howe, 2008). For this, the potential of large groups is harnessed. The crowd can have various motivations to contribute to the task, e.g., financial incentives, enjoyment, or social status (Yen, Dow, et al., 2016). Crowdsourcing is particularly useful for scaling complex work that cannot be handled by computers. As a result, crowdsourcing is being applied in various process areas of the software development process (Sarı et al., 2019). In the last years, the field of crowdsourcing in IS development and software engineering has grown quickly (K. Mao et al., 2017). The most used platform for crowdsourced software engineering is TopCoder (Sarı et al., 2019). TopCoder uses competitions to find the best solution and rewards the best participants with prize money. However, other crowdsourcing platforms that rely on collaboration instead of competition (e.g., Amazon Mechanical Turk (MTurk)) are also applicable for software development (Sarı et al., 2019). Besides the anonymous crowds that can be reached via dedicated platforms like MTurk or uTest, we consider for our review also crowds like employees, stakeholders, and students that are not necessarily recruited on these platforms. The IS development tasks to which crowdsourcing is mostly applied are requirements analysis, coding, and testing (Ambreen & Ikram, 2016; K. Mao

et al., 2017; Sarı et al., 2019). Applying crowdsourcing in IS design and development yields several advantages. For instance, TopCoder is reported to deliver software artifacts with a lower defect rate and higher quality at lower cost in less time compared to in-house development or outsourcing (Lakhani, Boudreau, et al., 2013; Lakhani, Garvin, et al., 2010). Common concerns regarding crowdsourcing in IS development involve intellectual property, quality, uncertainty, limited interaction, and collaboration overhead (K. Mao et al., 2017). There are already multiple reviews on crowdsourcing in IS development in general (K. Mao et al., 2017; Sarı et al., 2019) and on crowdsourcing in IS evaluation in particular (Alyahya, 2020; Leicht, 2018). However, so far, the focus has been put on crowdsourced software testing. Crowd-testing is an emerging trend in software engineering that enables companies to outsource different software testing activities to a large pool of workers (Alyahya, 2020). Dedicated platforms like uTest provide contact to thousands of workers and ensure that the individual testing requirements are included in the tasks (Alyahya, 2020). Crowd feedback has been considered only marginally, if at all, as an aspect of crowd-testing. However, crowd feedback goes beyond crowd-testing and should therefore be considered in a separate literature review. Additionally, a conceptualization is required to get a holistic understanding of existing approaches for crowdsourcing feedback within IS evaluation.

2.2.3. Conceptualization of Crowdsourcing Systems

Existing conceptualizations of crowdsourcing systems (e.g., Morschheuser et al., 2017; Pedersen et al., 2013; Zuchowski et al., 2016) all follow the same structure: First, a task or problem defined as "a statement of an initial condition and a desired ending condition" (Pedersen et al., 2013, p.581) is identified. This is followed by a specification of the crowdsourcing task and system, and finally, specific outcomes are achieved.

Pedersen et al. (2013) made the first effort to conceptualize crowdsourcing research in general. The first element of their conceptual model is the problem which defines the requirements for all other model elements. The main part of the model includes a process (the design of a step-by-step plan to solve the problem), governance (the actions and policies that are applied to manage the crowd), people (including the problem owner and the crowd that consists out of many individuals), and the technology (the technical capabilities that enable the formation of the crowd and the interaction and collaboration between

individuals in the crowd). The final element of the conceptualization is the outcome. This refers to the factual and the perceptual outcome of the crowdsourcing process. While the conceptualization of Pedersen et al. (2013) applies to crowdsourcing tasks in general, Zuchowski et al. (2016) and Morschheuser et al. (2017) focused on specific crowdsourcing tasks. The framework of Zuchowski et al. (2016) conceptualizes IT-enabled crowdsourcing with employees in enterprises, also called 'internal crowdsourcing'. This framework includes similar elements as the model of Pedersen et al. (2013). The main contribution of Zuchowski et al. (2016) is the definition of subdimensions that describe internal crowdsourcing tasks. Morschheuser et al. (2017) developed a conceptual framework for gamified crowdsourcing which is based on the previous two conceptualizations. As Morschheuser et al. (2017) put their focus on the crowdsourcing system this element replaces the main component which was initially defined by the governance, IT, process, and people. The gamified crowdsourcing system is mainly defined by the type of crowdsourced work. The design of the crowdsourcing system is not only influenced by the initial problem and tasks as in the other two frameworks but also by the crowds' motivation and the resulting behavior which is, in turn, a result of the gamification affordances and additional incentives. As we appreciate the approach of Morschheuser et al. (2017) of separating the crowdsourcing system and the crowd configuration, our framework will mainly be based on their framework of gamified crowdsourcing.

2.3. Research Methodology

To answer RQ1 and RQ2, we followed a multistep research approach. In the first step, we sought to review the current state-of-the-art and to derive a set of relevant papers on crowd-feedback systems. This set of papers provided our foundation to answer RQ1 and RQ2. In the second step, we used these papers to develop a conceptualization of crowd feedback in the form of a morphological box (RQ1). Finally, we used the set of papers as well as the morphological box and applied a cluster analysis to identify the existing research streams (RQ2). We outline each of these steps in more detail in the following sections.

2.3.1. Systematic Literature Review

To conduct the SLR we followed the guidelines of Kitchenham and Charters (2007) and first developed our search strategy. Therefore, we created the search string in several iterations.

We started with an exploratory search using Google Scholar with the search string "crowd AND feedback AND system". After reviewing the results, we iterated the search string several times using Google Scholar. The final search string consisted of four parts: The first part is for ensuring that a crowd is involved in the feedback collection process. In the second part, we added 'critique' and 'comment' as synonyms for feedback, and in the third part, we added 'method', 'process', and 'tool' as further means to crowdsource feedback. The fourth part was added to specify about what the feedback is collected. For searching for papers that collect feedback on IS, we additionally included specific types and characteristics of IS like 'website', 'software', 'interactive', 'app', and 'interface'. To provide a holistic overview we also included studies on crowdsourcing feedback on graphic and product design. Due to the similarity to user interface design, we expect the feedback systems and their features to be also applicable to feedback on user interface design. Therefore, we included the term 'design' in the fourth part of our search string. Finally, applying Boolean operators and wildcards led us to the final search string:

crowd* AND (feedback OR critique OR comment) AND (system OR process OR method OR tool) AND ("information system" OR website OR software OR interactive OR design OR app OR interface).

In the next step, we selected ACM Digital Library, IEEE Explore, and AISeL as databases for our SLR. These databases are well-established and were already used by scholars as reliable sources for literature reviews (Bandara et al., 2015). We decided to not limit the results to a specific time period or publication outlet in order to get a holistic overview. To refine the initial set, we then scanned the title, abstract, and keywords followed by reviewing the full text of the remaining papers. For the filtering, we applied six selection criteria: (1) the paper implements a prototype or develops a conceptual framework for collecting feedback, (2) the paper investigates an artifact mainly used to explicitly collect feedback, (3) the feedback is provided by a human crowd and assesses information systems, visual designs or product designs, (4) the article is peer-reviewed, (5) the article has more than three pages, (6) the article is written in English. Next, we conducted a backward and forward search following the same criteria. In our final set, we identified multiple articles that refer to the same crowd-feedback system. For the subsequent analysis of the papers, we inspect these papers jointly. Finally, we coded the resulting set of articles by their main research methodology. The main result of this step is a comprehensive set of papers

investigating crowd-feedback systems.

2.3.2. Concept Creation

In the next step, we analyzed the identified set of papers in order to conceptualize crowd-feedback systems (RQ1). Therefore, we developed a morphological box that captures all relevant dimensions of crowd-feedback systems. A morphological box provides a structured overview of all potential solutions to a problem (Zwicky & Wilson, 1967) and is commonly used for SLRs in the domain of IS to illustrate the diversity of solutions. To develop a morphological box, we followed the approaches of Wolfswinkel et al. (2013) and Nickerson et al. (2013). Based on their recommendations, we applied a three-step development approach:

In the first step, we followed Nickerson et al. (2013) and conducted a conceptual-to-empirical development approach to create an initial conceptual framework as a foundation for the following steps. Therefore, we started with the three frameworks introduced in the conceptual foundations from Morschheuser et al. (2017), Pedersen et al. (2013), and Zuchowski et al. (2016), extracted all of their dimensions and developed an initial conceptualization comprising of four overarching dimensions (i.e., input, crowd configuration, design characteristics, and effects) to guide our next steps.

In the second step, we again followed Nickerson et al. (2013) and conducted an empirical-to-conceptual development approach. Therefore, we used an inductive coding approach to create new subcategories for our morphological box and to identify codes for these subcategories based on Wolfswinkel et al. (2013). This step is necessary to develop a morphological box for crowd feedback accounting for their characteristics which are not captured by the initial coding scheme yet. For that, we iteratively reviewed each of the identified papers and continuously refined the initial coding scheme until the concepts reached an acceptable level of abstraction.

In the third step, all studies included in the final set were coded according to the concepts that we defined in the previous steps and a morphological box, as well as a concept matrix as described by Webster and Watson (2002), was created.

2.3.3. Cluster Analysis

Based on the identified set of papers and the derived morphological box, we were seeking to identify the existing research streams on crowd-feedback systems in order to answer RQ2. Thereby we aimed to understand which characteristics of crowd-feedback systems are usually combined and which effects are achieved by doing so. This shall help future researchers and practitioners to select appropriate design combinations when developing new crowd-feedback systems. Due to the relatively low number of papers, we first clustered the papers manually by identifying characteristics that often occur in combination and grouping these papers together. To verify the results, we decided to apply the two-step clustering analysis developed by Chiu et al. (2001). The two-step clustering is an effective approach to identifying clusters and is often applied in literature reviews (e.g., Knaeble et al., 2020; Rissler et al., 2017). The advantage of this approach compared to pure hierarchical clustering is that it automatically detects the optimal number of clusters and provides the silhouette measure of cohesion and separation as a quality measure. Additionally, it summarizes the influence of each characteristic on the cluster allocation which helped us to verify if we identified the correct characteristics as the main drivers of the clustering. To conduct the two-step clustering, we first transformed our concept matrix into a binary form by changing every 'X' to a '1' and every empty cell to a '0'. Then we applied the two-step clustering to separate our articles into homogenous groups (clusters) using IBM SPSS Statistics 27. The two-step clustering is based on two distinct steps: First, the entire dataset is scanned, and based on sequential clustering preclusters are created. In this step, the log-likelihood distance measure is applied as the similarity criterion which is appropriate as our input data is binary (Sarstedt & Mooi, 2014; Theodoridis & Koutroumbas, 2009). Second, agglomerative hierarchical clustering is applied to the created preclusters (Chiu et al., 2001). We applied Akaike's information criterium (AIC) to determine the appropriate number of clusters (Akaike, 1998).

2.4. Results

In this section, we describe the results of our three-step research approach. First, we outline the results of the SLR and describe the identified set of papers. This serves as the foundation for the following two steps. Second, we present the morphological box of crowd-feedback systems based on the discovered articles and the iterative refinement of the

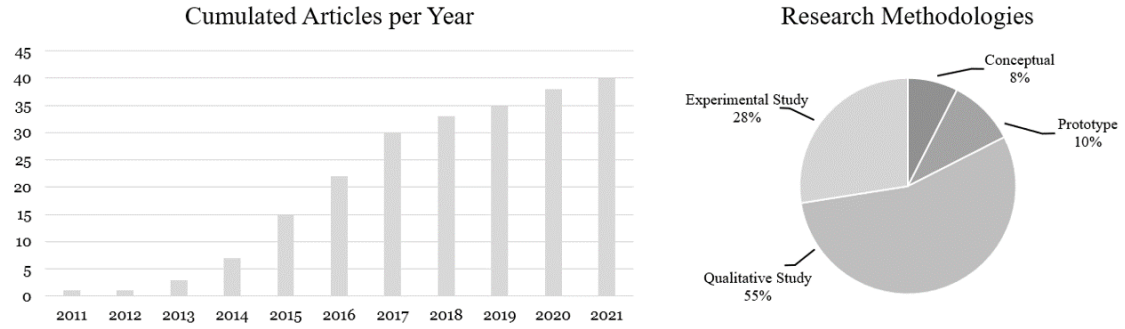


Figure 2.1.: Cumulated Number of Articles reporting Crowd-Feedback Systems (left) and applied Research Methodologies (right)

coding dimensions (RQ1). Third, we introduce three research streams that were identified via the cluster analysis (RQ2).

2.4.1. Results of Systematic Literature Review

Applying our search string to the identified databases (i.e., AISeL, ACM Digital Library, IEEE Explorer) resulted in 274 studies. From the initial 274 studies, we excluded 227 by carefully scanning title, abstract, and keywords and thereby applied our inclusion criteria (47 remained). We applied the same criteria when reviewing the full texts and kept 24 studies. Most of the excluded studies either focus on crowd-testing and address feedback only marginally, or collect feedback on something else than information systems, visual designs, or product designs (e.g., university courses). Finally, we conducted a backward and forward search following our criteria and identified 16 additional studies. Consequently, in total, we identified 40 relevant articles. Since some of these studies refer to the same system, only 34 different crowd-feedback systems are included in our set of studies.

For the descriptive information on our paper set, we considered all 40 articles without excluding articles on the same crowd-feedback system. This is necessary to account for a holistic overview of all existing studies. A complete list of all identified papers is depicted in Appendix Table A.1. The analysis of the publication dates of the articles (see Figure 2.1, left) shows that the topic of crowdsourcing feedback emerged around ten years ago and had a peak between 2015 and 2017. The most common research methodology applied is the qualitative study including case studies as well as grounded theory (see Figure 2.1, right). It is followed by the experimental study methodology which is applied in 28% of the studies. Our set also includes articles on prototype development and conceptual models for crowd-feedback systems.

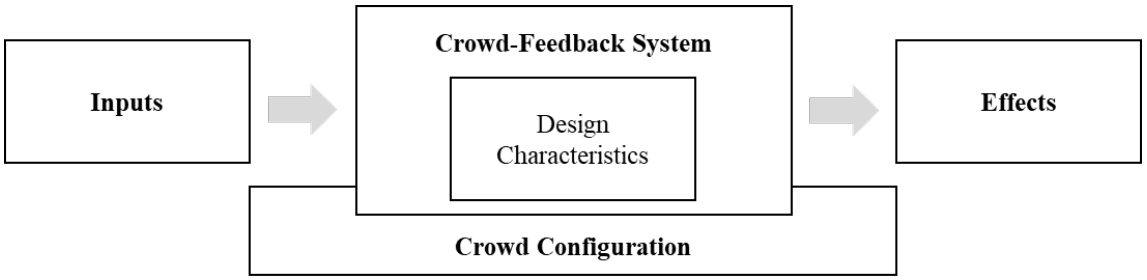


Figure 2.2.: Conceptual Framework of Crowd Feedback

2.4.2. Results of Concept Creation

In the first step, we developed an initial conceptual framework (see Figure 2.2) based on the existing conceptualizations of crowdsourcing tasks and systems (Morschheuser et al., 2017; Pedersen et al., 2013; Zuchowski et al., 2016) and the identified 40 papers to describe the existing research on crowd feedback. We retained the established components 'problem' and 'outcomes' and renamed them 'input' and 'effects'. In our case, the input includes characteristics of the IS on which the feedback shall be collected and the specification of the feedback that is requested. The effects describe the effects of applying crowd-feedback systems on the crowd, the feedback, and the resulting IS design. The crowd-feedback system that we put in the center of our conceptualization is mainly described by its design. The crowd-feedback system is usually intertwined with the crowd configuration as the design of the system might restrict the crowd configuration and vice versa.

In the next step, we iteratively developed subdimensions and concepts for each of the four dimensions of our initial conceptual framework to capture the complete landscape of crowd-feedback systems. The development of the subdimensions was completely based on the 40 papers of our SLR. The subsequent coding of all papers resulted in a concept matrix (see Appendix Table A.1) and a morphological box (see Figure 2.3) that includes all (sub)dimensions and characteristics. Studies may include several characteristics of one subdimension (e.g., they collect qualitative and quantitative feedback). For each concept, the figure shows the absolute frequency of the concept in our set of papers (indicated by the number behind the concept). The different shades visualize this frequency and help to illustrate the current focus of research. In the following, we provide an analysis of each subdimension.

Dimension	Subdimension	Characteristic				
Inputs	IS Lifecycle Stage	Development (28)			Operations (10)	
	Feedback Type	Qualitative (29)			Quantitative (17)	
	Scope of Feedback	Non-functional Attributes (25)		Content (17)		Functional Attributes (15)
Crowd Configuration	Crowd Type	Annoymous (21)		'Proxy' Users (13)	Students (3)	Convenience (2)
	Incentive	Money (22)	Involvement and Improvement (8)	Interest and Social Compensation (5)	Credits (3)	Gamification (2)
Design Characteristics	Feedback Collection Mechanisms	Questionnaire (10)	Free Text Field (8)	Categories (8)	Selection (7)	Direct Manipulation (2)
	Interactivity Cues	Collaboration (7)		Marker (7)	Context (4)	Recording (4)
Results		Outcome Effects (22)		Process Effects (16)		Intermediate Effects (7)

Applied in ... of studies: < 25% < 50% < 75% ≥ 75%

Figure 2.3.: Morphological Box for Crowd Feedback

2.4.2.1. Dimension: Inputs

The input dimension describes characteristics of the initial situation of the feedback collection process. Thereby, we distinguish between characteristics of the IS that need to be evaluated, like its lifecycle stage, and the characteristics of the feedback that is sought. While the type of feedback indicates if the collected feedback is of a quantitative or qualitative nature, the feedback scope represents the attributes of an IS that the feedback is focused on. Some of the crowd-feedback systems are developed for collecting feedback on one specific IS like an 'interactive Energy Saving Account' (Stade et al., 2017) or e-services of public administrations (Pretel et al., 2017). Others focus on an IS class like conversational user interfaces (Choi et al., 2021; Yu et al., 2016), mobile apps (Ayalon & Toch, 2018, 2019; Oppenlaender, Kuosmanen, et al., 2021; Seyff, Ollmann, & Bortenschlager, 2014), reinforcement learning systems (de la Cruz et al., 2015) or adaptive software systems (Muñante et al., 2017). However, most of the presented crowd-feedback systems are not limited and should apply to a wide range of IS.

IS Lifecycle Stage. For the IS lifecycle stage, we distinguish between systems that collect feedback for IS during its development process and systems that collect the feedback during operations of the IS for further improvement. Most of the articles (28) cover collecting feedback during the development stage, while only ten systems collect feedback during operations. Only four of the feedback systems apply to both lifecycle stages. For instance, Snijders et al. (2015) built a gamified online platform that can be applied to elicit requirements for new IS as well as requirements to further improve existing IS. Y. W. Wu and Bailey (2016) evaluated their system by collecting feedback on an existing web page.

However, their crowd-feedback system is designed in a way that is also applicable to evaluating IS during development.

Feedback Type. Most papers (29) seek qualitative feedback which is usually done via text fields. However, quantitative feedback is also often collected and is requested in various ways. A common way is to vote designs 'up' or 'down' as it is applied in the *CrowdUI* system (Oppenlaender, Tiropanis, & Hosio, 2020). There, participants can vote on the design suggestions of other participants that were created by manipulating the initial design. The system of Jansson and Bremdal (2018) asks users to vote on web page elements to help a system based on artificial intelligence to learn user preferences. Easterday et al. (2017) implemented a feedback system similar to a forum where users can add qualitative comments, but also vote on the comments of other participants. Besides the voting, the selection of items is a common method to collect quantitative feedback. By asking users to select abstract and emotional images that represent their emotional reaction to a particular design, Robb, Padilla, Kalkreuter, and Chantler (2015a, 2015b) and Robb, Padilla, Methven, et al. (2017) introduced an innovative way to collect quantitative feedback. *Paragon* (Kang et al., 2018) enables participants to enrich their feedback by selecting exemplary designs. Finally, the most common way to collect quantitative feedback are Likert scales as used by Schneider et al. (2016) to indicate how severe a usability problem is or by Oppenlaender, Kuosmanen, et al. (2021) and Xu and Bailey (2014) to rate if design guidelines are considered.

Scope of Feedback. Feedback can be collected on non-functional attributes, thus aesthetics and human values, functional attributes that summarize feedback about features and functionalities of the IS or content of the IS. The code content is meant for systems that collect feedback about the information that is provided by the IS, for example, the content of a website. While non-functional attributes are considered in 25 of our 34 studies, feedback on functional attributes and content is less often included. As many of the papers in our set focus on design feedback, they usually ask only for feedback on visual design and aesthetical aspects like layout, consistency, balance, readability, and simplicity (Luther, Pavel, et al., 2014) or aim to understand the first notice and impressions of viewers (Xu & Bailey, 2014). Besides these aspects, users' human values are another non-functional attribute. Here, crowd-feedback systems are used to evaluate if the respective values are considered in the IS design. While Ayalon and Toch (2018, 2019) examine the social and

institutional privacy of IS, (Hosseini et al., 2016) developed a system to collect feedback on the implementation of transparency requirements. Feedback on functional attributes focuses on the capabilities of the IS. *Citizenpedia* (Pretel et al., 2017) is a platform for citizens to comment on current procedures of e-services. For this purpose, the platform provides a hierarchical overview of the services offered and the corresponding flow of interactions between the citizens and the public administration. There exist also two studies on collecting feedback on the functionality of conversational user interfaces (Choi et al., 2021; Yuan et al., 2016). Both tools focus on providing feedback on answers of the conversational user interface. Yu et al. (2016) ask participants to rate the appropriateness of chatbot reactions. Choi et al. (2021) change the conversation flow or provide new suggestions for chatbot reactions. Feedback on content is only used as an addition to feedback on functional or non-functional attributes. None of the articles in our set collected feedback only on the content. The collection of feedback on the content of the IS is especially useful for IS that aim to provide information like a university website (Y. W. Wu & Bailey, 2016), posters and logos (Yen, Dow, et al., 2016), or weather dashboards (Krause et al., 2017).

2.4.2.2. Dimension: Crowd Configuration

The crowd configuration describes how the crowdsourcing task is configured by analyzing which type of crowd is asked for feedback and how this crowd is incentivized to contribute.

Type of Crowd.

The type of crowd can be anonymous when dedicated crowdsourcing platforms are used or can consist of 'proxy' users, students, or friends and family (coded as convenience). Most of the papers in our set use an anonymous crowd to collect feedback. These crowds are recruited on platforms like MTurk, Mobile-Works, or Upwork (Greenberg et al., 2015; Krause et al., 2017). Besides crowdworkers that were recruited on dedicated crowdsourcing platforms, 'proxy' users are also frequently used for feedback collection. This term includes actual and potential users as well as other stakeholders like developers, analysts, clients, and regulatory bodies (Snijders et al., 2015). Students are only used as feedback providers when the IS design is part of an educational class as in the studies of Oppenlaender, Kuosmanen, et al. (2021) and Robb, Padilla, Kalkreuter, and Chantler (2015a). In the studies of Wauck et al. (2017) and Yen, Dow, et al. (2016), designers use social media platforms to crowdsource feedback from social contacts. In both studies, convenience

feedback is not used as the main source of feedback but to compare the feedback with the feedback from other sources. Another way to involve contacts from social networks in the feedback process is to promote the feedback system on these platforms as done by Haukipuro et al. (2016).

Incentives.

For incentives, we distinguish between money, involvement in the IS development and improvement of the IS, interest and social compensation, course credits, and gamification. The incentives are highly related to the type of crowd used. While the anonymous crowd is usually incentivized by a financial reward, 'proxy' users are motivated by the prospect of improvement of the IS and the feeling of being involved in the development process. Students are usually incentivized by course credits and social contacts contribute because of personal interest and social compensation. Gamification is the only incentive that we identified that is not related to one type of crowd. Although Pretel et al. (2017) and Snijders et al. (2015) apply gamification only in combination with actual user feedback, Morschheuser et al. (2017) show that gamification can also be applied to motivate other types of crowds.

2.4.2.3. Dimension: Design Characteristics

The design characteristics describe the features of the proposed crowd-feedback systems. During the analysis of our set of articles, we learned that crowd-feedback systems usually consist out of a feedback collection mechanism and some additional functionalities that aim to support the crowd during the process of providing feedback, here called interactivity cues. In this section, we analyze both parts of crowd-feedback systems separately.

Feedback Collection Mechanisms. For the feedback collection we distinguish between five mechanisms: With questionnaires, participants are asked a series of questions. These can be of a qualitative or quantitative nature. Categories enable participants to select a suitable category or rubric for their feedback. Compared to questionnaires, this mechanism offers more freedom to feedback providers. Systems that provide only one single text field for feedback, are coded with free text field. Another mechanism is the selection, where the feedback is provided by selecting items. The most complex mechanism to provide feedback is direct manipulation where the crowd can edit a system according

to their needs and wishes. The frequency of occurrences of the five feedback collection mechanisms that we identified is evenly distributed. Only direct manipulation is less used. Questionnaires include either closed questions (Haukipuro et al., 2016; Nebeling et al., 2013) or open questions (Greenberg et al., 2015; Xu & Bailey, 2011). Using questionnaires can be advantageous compared to categories or more rigid structures, especially when dealing with a large range of themes and possible critiques (Greenberg et al., 2015). On the other hand, categories offer more freedom for the participants, since they can often specify several feedback points under one category and can also choose the order that they want to provide feedback themselves (Luther, Pavel, et al., 2014; Yuan et al., 2016). The study of Yuan et al. (2016) shows that rubrics help feedback providers to contribute feedback that is nearly as valuable as expert feedback. They show that feedback collection via rubrics leads to feedback with higher quality than feedback collected in free text fields. Feedback systems that collect feedback via a free text field do not ask the users specific questions but just provide a field to enter feedback (Krause et al., 2017; Seyff, Ollmann, & Bortenschlager, 2014; Y. W. Wu & Bailey, 2021). This mechanism is also often used in user forums and similar communities like the gamified platform to derive user requirements of Snijders et al. (2015). A very different way to collect feedback is selecting items, e.g., design elements (Jansson & Bremdal, 2018), concrete improvements (Choi et al., 2021; de la Cruz et al., 2015), or labels for chatbot answers (Yu et al., 2016). The selection mechanism enables the collected data to be easily quantifiable and may therefore reduce the effort to analyze the feedback. Finally, direct manipulation is the most direct way for the crowd to provide feedback. Oppenlaender, Tiropanis, and Hosio (2020) developed a feedback system that enables the crowd to change the UI design of an existing IS and the *ProtoChat* system (Choi et al., 2021) allows the user to change the conversation flow of the chatbot.

Interactivity Cues. We identified four frequently used interactivity cues: collaboration, marker, context, and recording. There exist more than these four, but we chose to focus on cues that appeared in more than one single article. Collaboration means that the participants can interact with the feedback of other crowd members by rating or commenting on it. Systems that enable the crowd to mark their feedback visually in the system are coded with marker. Context means that the crowd receives a context of use in the form of a specific scenario or a persona before providing their feedback. Recording is for systems

that include a feature that allows participants to do voice or video recordings of their feedback. Collaboration is often applied in crowdsourcing systems that resemble a user forum or community. There, users can see the feedback of others and react to it by commenting and voting (Haukipuro et al., 2016; Pretel et al., 2017; Snijders et al., 2015) or just use it as inspiration (Xu & Bailey, 2011). Another way to include collaboration is to let the crowd rate design suggestions of other crowd members (Oppenlaender, Tiropanis, & Hosio, 2020). Markers can either be in the form of flags that can be put onto the IS user interface to indicate what specific element is meant by the feedback (Luther, Pavel, et al., 2014; Y. W. Wu & Bailey, 2016; Xu & Bailey, 2014) or in the form of screenshots and photos that are taken of the IS when the user encounters a problem and wants to provide feedback (Schneider et al., 2016; Seyff, Ollmann, & Bortenschlager, 2014; Stade et al., 2017). Oppenlaender and Hosio (2019) include the marker feature in their system as the actual answer to tasks like 'Touch the areas of the artwork that you like best!'. The context of use aims to help the participant to imagine himself in a real usage scenario while providing feedback. The context can either be created by providing a persona (Ayalon & Toch, 2018; Muñante et al., 2017) or a scenario that explains the design and its context (Ayalon & Toch, 2019; Y. W. Wu & Bailey, 2021). Recording feedback is an optimal way to better capture the emotions of feedback providers (Ma et al., 2015) and reduce the effort for them to provide feedback (Seyff, Ollmann, & Bortenschlager, 2014). While Ma et al. (2015), Easterday et al. (2017), and Dow et al. (2013) let participants record videos of their feedback, Seyff, Ollmann, and Bortenschlager (2014) enable users to provide short audio recordings of their feedback.

2.4.2.4. Dimension: Effects

Finally, this dimension specifies the consequences of applying the described crowd-feedback system. For the effects, we did not define subdimensions, but three codes: process effects, intermediate effects, and outcome effects. As the studies usually only investigate a subset of possible effects, we can only report the effects that are described in the studies although the feedback system might lead to further effects as well. Process effects address the well-known problems of user-centered evaluation methods like scalability (de la Cruz et al., 2015; Easterday et al., 2017; Oppenlaender, Tiropanis, & Hosio, 2020; Schneider et al., 2016; Yen, Dow, et al., 2016), effort (Ayalon & Toch, 2019; Greenberg et al., 2015; Haukipuro et al., 2016; Hosseini et al., 2016; Jansson & Bremdal, 2018; Ma et al., 2015) and costs

(de la Cruz et al., 2015; Greenberg et al., 2015; Ma et al., 2015; Schneider et al., 2016) of the feedback collection process. Additionally, the diversity of the feedback providers and the resulting feedback is often seen as one big advantage of applying crowdsourcing (Dow et al., 2013; Haukipuro et al., 2016; Ma et al., 2015; Nebeling et al., 2013; Oppenlaender, Kuosmanen, et al., 2021; Schneider et al., 2016; Wauck et al., 2017). These outcomes are in most cases elicited via qualitative interviews with the respective feedback requesters. Intermediate effects are outcomes that might mediate outcome effects like the feedback quality but are not the direct goal of applying crowd-based feedback systems. These effects include user engagement (Hosseini et al., 2016; Oppenlaender & Hosio, 2019; Robb, Padilla, Kalkreuter, & Chantler, 2015b; Robb, Padilla, Methven, et al., 2017; Snijders et al., 2015; Yu et al., 2016) and inspiration for the designer (Jansson & Bremdal, 2018; Kang et al., 2018; Robb, Padilla, Kalkreuter, & Chantler, 2015a, 2015b). Finally, the outcome effects are usually measured via a quantitative evaluation, often including experts or feedback requesters to rate the quality and helpfulness of the feedback or the final IS designs. The most mentioned outcome of crowd-feedback systems is the quality of feedback, including reliability and helpfulness of feedback (e.g., Ayalon and Toch, 2018 Choi et al., 2021; Krause et al., 2017; Luther, Pavel, et al., 2014). Additional aspects of outcome effects are the quantity of feedback (Kang et al., 2018), a better final design (Lekschas et al., 2021; Luther, Tolentino, et al., 2015; Xu, Rao, et al., 2015), and increased user satisfaction when using the improved IS (Muñante et al., 2017).

2.4.3. Results of Cluster Analysis

By manually grouping papers with similar characteristics, we identified three clusters. Thereby, we identified the input and crowd configuration, especially the feedback scope and crowd type, as the main drivers for the cluster affiliation. The two-step cluster analysis confirmed our assumption and identified similar research streams while the optimal number of clusters was found to be three. The most important categories are the crowd type, the feedback scope, and the incentive which is highly related to the crowd type. The research streams based on the two-step clustering are displayed in Appendix Table A.1. The silhouette measure of cohesion and separation of the analysis is 0.3 which indicates a medium solution (Sarstedt & Mooi, 2014). However, since we obtained similar results from the manual analysis, we consider the results reliable and present them in the following. Thereby we highlight the characteristics of each research stream to complement the

comprehensive state-of-the-art overview provided in the previous chapter.

Stream 1 – Anonymous Crowd Feedback: The first stream (11 studies) is dominated by crowd-feedback systems that are designed to ask an anonymous crowd for feedback. These studies mostly originate from the field of visual design. Here, the crowd is usually incentivized by money. Feedback is collected during the development stage on non-functional attributes and often additionally on content. Thereby, all studies collect qualitative feedback. Consequently, these systems are mainly focused on formative feedback to further improve the design. In these studies, the focus lies on achieving outcome effects.

Stream 2 – Real User Crowd Feedback: The 10 studies in this stream are mainly performed to collect feedback from real and potential users on systems during development as well as systems in operations. The crowd is incentivized by involvement and improvement of the system and is asked for qualitative and quantitative feedback mostly on functional attributes. Most of these systems apply collaboration as an interactivity cue. Therefore, these crowd-feedback systems resemble user forums but provide more guidance. The outcomes of user forums are in most cases outcome effects and sometimes process effects.

Stream 3 – Hybrid Crowd Feedback: This stream includes 13 studies and is less clearly defined than the other two streams. In this stream, the studies ask all types of crowds for feedback with no limitation on specific attributes. The studies are mainly connected by the goal of achieving process effects. This is consistent with the fact that feedback is mostly collected through questionnaires and selection. Besides process effects, some studies additionally achieve intermediate effects like increased user engagement.

2.5. Discussion

This study synthesizes characteristics of crowd-feedback systems for IS development from articles reporting the results of research projects in this field. Regarding the conceptualization of crowd feedback (RQ1), we developed a conceptual framework and a morphological box for crowd feedback. These conceptualizations are not limited to the configuration of crowd-feedback systems but include associated aspects like crowd configuration as well. Our morphological box provides a comprehensive structure of crowd-feedback systems and visualizes where the focus in recent research was put. The morphological box can be applied to future research projects to consider possible design choices. The subsequent

cluster analysis helped us to identify patterns in the existing studies and showed which characteristics of crowd feedback are usually connected. This might also support crowd-feedback system developers in making the right design choices considering their specific use case. Based on the insights we gained during our study, we now propose four main future research directions (RQ2).

2.5.1. Effects of Design Characteristics

First, we argue that there is a need to better understand the effects of specific design characteristics. Our analysis revealed several feedback collection mechanisms and interactivity cues that can be included in crowd-feedback systems. By clustering our papers, we identified a connection between the inputs, the crowd configuration, and the resulting effects. However, we could not find a pattern for most of the design characteristics. While questionnaires and selection are related to process effects as they are easily quantifiable and therefore scalable, the other feedback collection mechanisms do not seem to be connected to inputs, crowd configuration, or effects. The same applies to the interactivity cues. Collaboration occurs mostly when real users are asked for feedback, but the remaining interactivity cues seem to not follow any pattern. Although there exists already some research that investigates specific design characteristics, such as the effect of rubrics on the feedback quality (Yuan et al., 2016), there is still a lack of systematic research on the design characteristics of crowd-feedback systems. For the application of crowd-feedback systems in practice, it is important to understand how to achieve specific effects by selecting the appropriate design characteristics, not only considering the feedback quality and quantity but also the effects on the crowd. At the same time, we consider it important to learn how the design characteristics are related to the input characteristics and crowd configurations to provide recommendations according to the selected inputs. While we assume most interactivity cues have a positive impact on the feedback, there might also be some drawbacks. For example, we would assume that a collaboration feature positively impacts the provided feedback. However, the crowd might either be inspired by the comments of others or could be influenced by the perceptions of *opinion leaders* (Bodendorf & Kaiser, 2009) and consequently not share their own opinions. They could even get the feeling that their feedback is not required anymore. To investigate individual effects as well as the effects of combining characteristics, dedicated experimental studies are required.

2.5.2. Intermediate Effects on the Crowd

Second, our concept matrix shows that intermediate effects are only investigated by a few crowd-feedback studies. In these studies, only user engagement and the inspiration of the designer were investigated as intermediate effects. Additionally, the connection between the interactivity and the intermediate effects, especially concerning the crowd's behavior and perceptions of the crowd-feedback system, needs to be further investigated. Related studies in the field of IS development already identified a positive influence of interactive features on user engagement and the resulting behavior for employee participation (Feine et al., 2020). According to the insights of this study, intermediate effects could also serve as mediators for outcome effects like feedback quantity and quality. Besides user engagement, we suggest exploring additional intermediate effects such as the perceived interactivity, the effort of using the crowd-feedback system, as well as its usability. Knowledge of the connection between these constructs, design characteristics, and the resulting feedback will help feedback requesters to design better feedback systems.

2.5.3. Crowd-Feedback System Configurators

Third, we propose to research crowd-feedback system configurators to enable novices to build and adapt crowd-feedback systems according to their individual use case. All existing crowd-feedback systems consist out of a fixed set of design characteristics and provide no functionality to adapt them to a specific use case. The configurator should be based on the results of the two previously suggested avenues for future research and consider the three clusters that we identified in this paper. Adaptable crowd-feedback systems might not only make crowd feedback applicable to a more diverse set of use cases (Luther, Tolentino, et al., 2015) but also increase the feedback quality and user satisfaction (Almaliki et al., 2014).

2.5.4. Continuous Feedback Collection

Fourth, we identified the need to research further support for continuous IS evaluation. As we highlighted in the beginning, continuous user involvement is crucial for IS acceptance and success (Harris & Weistroffer, 2009; Ives & Olson, 1984). During the analysis, we learned that most studies that investigate crowd feedback focus on feedback during the development process. Only four studies in our set developed a system that applies to the

feedback collection during the development and operations of the IS. However, none of these studies explicitly decided to focus on the entire lifecycle, but the respective crowd-feedback systems can just be used to evaluate IS during development or operations. To ensure continuous user involvement a crowd-feedback system that supports the development team during the whole lifecycle is necessary. This system could be combined with the crowd-feedback system configurator. Thereby, this system should guide the researcher in adapting the crowd-feedback systems' features to the specific requirements of the context and the lifecycle stage of the IS.

2.5.5. Limitations

We are aware that our literature review has limitations. Firstly, our results are highly dependent on the search string, the selected databases, and the chosen selection criteria. Our selections may induce a bias in the extracted literature and impact the identified research streams. The high number of studies that we identified via the backward and forward citations shows that our search string had some shortcomings. However, to reduce the probability of bias, we applied established methodological recommendations (i.e., Kitchenham and Charters, 2007; Webster and Watson, 2002). All decisions during the three stages of our literature review are made explicit. Secondly, for the interactivity cues, we considered only the most common ones. This restricts the holistic overview we aimed to provide of crowd-feedback systems as well as it might influence the clustering. Thirdly, we are aware that the cluster quality of the three research streams that we identified is rather low. Nevertheless, we decided to report the three research streams as they were consistent with the results of the manual clustering. Additionally, we want to guide future researchers who can revise the clusters in further studies.

2.6. Conclusion

Besides crowd testing, crowd feedback is a promising approach to scale the continuous evaluation of IS. As current research lacks a comprehensive overview of existing crowd-feedback systems, we aimed to structure and analyze existing literature with three main contributions: First, we provided an overview of existing crowd-feedback systems by conducting an SLR and identifying 40 relevant papers. Second, we proposed a morphological box for structuring crowd feedback in IS development. Third, we identified three main

research streams for crowd-feedback systems. Based on the insights gained by these three contributions, we finally highlighted avenues for feature research. We believe that our SLR can serve as a reference in the broader field of crowd-feedback systems and the dimensions that should be considered when researching such systems.

3. Study II: Aligning Crowdworker Perspectives and Feedback Outcomes in Crowd-Feedback System Design

3.1. Introduction

The continuous integration of potential users in the evaluation of software is a challenging but critical activity in the design and development process (Brhel et al., 2015). However, due to their face-to-face character, traditional evaluation methods such as interviews, focus groups, or usability tests lack scalability and are costly. Furthermore, as they are usually conducted with small groups of participants, evaluation results tend to be limited concerning generalizability (Mackay, 2004).

Leveraging crowdsourcing in software development has received growing attention in research and practice. Commercial platforms like UserTesting, uTest, UserZoom, and UserCrowd offer different forms of crowdsourced evaluation services. In recent years, two research streams have emerged that have the goal to overcome the limitations of traditional software design evaluation forms through crowdsourcing: crowd testing and crowd feedback. Both focus on using the crowd to involve users in software development but differ in their objectives. Crowd testing has the goal of identifying system errors and follows existing testing methods like usability testing (Leicht, 2018). Crowd feedback aims to collect individual opinions and perceptions of the software design by users, anonymous crowdworkers, students, or friends and family (Haug & Maedche, 2021a). It is rooted in the field of visual design where peer feedback is an established approach to iterate design solutions (Wauck et al., 2017). Since it is not required for crowd feedback to have a high-fidelity prototype, but user stories or screenshots are sufficient, the application of crowd feedback is broader and more flexible. Moreover, crowd feedback is applicable throughout the whole software lifecycle and enables designers to collect diverse feedback in terms of type and scope (Haug & Maedche, 2021a).

Previous research proposed crowd-feedback systems that include various design features and can be applied in a diverse set of contexts. One of the most popular systems is

CrowdCrit (Luther, Pavel, et al., 2014). *CrowdCrit* mainly relies on qualitative feedback that users can add to predefined feedback categories. Additionally, users can apply markers to indicate which element or area their feedback is addressing. Thereby, *CrowdCrit* is mainly designed to evaluate static designs, like posters. There exist only a few crowd-feedback systems that focus on evaluating interactive design prototypes or even software, like *AppEcho* (Seyff, Ollmann, & Bortenschlager, 2014), *Critiki* (Greenberg et al., 2015), and *CrowdUI* (Oppenlaender, Tiropanis, & Hosio, 2020).

The majority of existing studies focus on demonstrating the feasibility of crowdsourcing feedback in their individual area of application. Thereby, mainly qualitative evaluation has been performed. Only a few studies have investigated the effects of design characteristics of crowd-feedback systems on the feedback quality and quantity in experimental studies following a quantitative evaluation approach. For example, Yuan et al. (2016) showed that offering novice crowdworkers feedback categories to indicate on which topics feedback is required has a positive impact on the feedback quality. Other studies showed the effects of very specific characteristics and requirements of the feedback like using a critique style guide (Krause et al., 2017), framing feedback as questions (Lekschas et al., 2021), or viewing the design on which the feedback shall be collected as part of a narrative (Y. W. Wu & Bailey, 2021). However, there exist many different design features of crowd-feedback systems that are frequently applied. These include, but are not limited to questionnaires, free text fields, categories, selection, direct manipulation, recordings, collaboration, markers, and scenarios (Haug & Maedche, 2021a). However, their individual effects on feedback quality and quantity are not well understood. This represents an important first research gap for the design of crowd-feedback systems.

Additionally, existing studies mainly focused on understanding the requirements of feedback requesters (i.e., designers), but fail to consider the perspective of feedback providers (i.e., crowdworkers). Oppenlaender and Hosio (2019) addressed this issue by comparing feedback providers' and requesters' feature preferences. However, their evaluation did not study the underlying reasons for users' preferences and did not analyze the resulting feedback outcomes. Additionally, not all insights can be transferred to the evaluation of interactive design prototypes or even software. Krause et al. (2017) also included crowdworkers in the evaluation of their critique style guide. Still, the crowdworkers' perspective represents only a minor part of the entire evaluation study. We believe that it is important

to include the perspective of feedback providers not only in the evaluation but also in the initial design of crowd-feedback systems. This allows us to align crowdworkers' requirements with the feedback outcomes. Understanding the effects of individual design features on the feedback and the feedback provider will help to adapt crowd-feedback systems better to their context of use. Thus, designers may be supported in selecting the appropriate design features considering their individual situations. This, in turn, will enable feedback requesters to apply crowd-feedback systems and help make the software development process not only more scalable but also even more human-centered. We identify this as a second major research gap in the field of crowd-feedback systems.

In this paper, we address these research gaps with two studies. In the first study, we conducted initial exploratory interviews to better understand the requirements of feedback providers. We explored how feedback providers perceive crowd-feedback system features and understood how these features should be implemented. Based on these insights, we developed *Feeasy*, a crowd-feedback system (Haug & Maedche, 2021b). *Feeasy* includes five key features: (1) a description of a usage scenario of the underlying design prototype to offer feedback providers a context, (2) a speech-to-text feature to add feedback comments via voice, (3) a marker feature to specify the elements of the prototype which the feedback addresses, (4) feedback categories to allocate the feedback comment to a specific category, and (5) a star rating for each category to collect additional quantitative feedback. We, subsequently, conducted an experimental study with *Feeasy* as an experimental artifact that analyzes the effects of crowd-feedback systems with different design features on feedback quality, quantity, and crowdworker perceptions. The feedback quality is measured via the assessment of UI-design skilled crowdworkers who evaluate each feedback comment in five quality categories (helpfulness, specificity, relevance, sentiment, and objectivity). The feedback quantity is measured via the length of feedback comments. In this study, we applied seven treatment conditions, one for each design feature, one basic treatment with no design features, and one full treatment with all five features combined. To further enhance our understanding of the crowdworkers' perspective on crowd-feedback system features, we conducted additional semi-structured interviews. Our results provide evidence that more design features are not beneficial in all use cases, but applying any design features is better than none. Furthermore, we learned that overwhelming feedback providers might

reduce feedback quality and quantity and that scenarios are the favorable design feature when considering the crowdworkers' perspective. With our results, we contribute and extend previous research on crowd-based user involvement in the software development process by analyzing and synthesizing the effects of five crowd-feedback design features and thereby aligning crowdworkers' perceptions with feedback outcomes. Thereby, we aim to allow future crowd-feedback systems not only to be more efficient and effective but also to improve the feedback experience for feedback providers (e.g., crowdworkers).

3.2. Conceptual Foundations & Related Work

3.2.1. User Evaluation Methods

Prominent methods to evaluate software designs with users are interviews, focus groups, and usability tests (Gibbs, 1997; Vredenburg et al., 2002). In general, these methods have in common that the involved designers, domain experts, and end-users have to meet virtually or physically to conduct the software usability and user experience (UX) evaluation. Consequently, these methods lack scalability, are time-consuming, and require monetary resources (Gibbs, 1997; Scholtz, 2001). One solution for these challenges is leveraging crowdsourcing. Specifically, dedicated crowdsourcing platforms are used to evaluate software design solutions (Haug & Maedche, 2021a). Crowdsourcing increases the scalability of software evaluation and reduces the effort for software developers and designers through its low-barrier accessibility (Ayalon & Toch, 2019; Greenberg et al., 2015). Additionally, it provides access to a diverse group of people to evaluate the software design (Ma et al., 2015). As introduced earlier, the application of crowdsourcing for evaluation purposes can be distinguished between crowd testing and crowd feedback (Haug & Maedche, 2021a). While crowd testing requests the crowd to conduct tests to identify errors in a system, crowd feedback asks users for their verbal feedback that includes opinions on and perceptions of a system. Therefore, crowd feedback may be conducted on interactive prototypes, static designs like screenshots and wireframes, and even textual descriptions like user stories. Crowd testing, in turn, requires high-fidelity prototypes that allow for interaction and include the original content of the system. In summary, crowd feedback allows us to intuitively evaluate the entire software design process from user stories to high-fidelity prototypes, and is well suited for this application.

3.2.2. Crowd-Feedback Systems

There exist multiple systems that support software designers and developers in collecting design feedback on crowdsourcing platforms. Crowd-feedback systems differentiate in the form of multiple dimensions (Haug & Maedche, 2021a). With regards to the subject under investigation, recent crowd-feedback systems focus on collecting feedback on visual designs such as posters (Luther, Pavel, et al., 2014), specific software applications such as chatbots (Choi et al., 2021), and websites (Oppenlaender, Tiropanis, & Hosio, 2020), or mobile apps (Seyff, Ollmann, & Bortenschlager, 2014). Thereby, the systems differ in the phase of the development lifecycle they are focusing on. While some systems focus on collecting feedback during the development process (e.g., Oppenlaender, Kuosmanen, et al., 2021; Schneider et al., 2016; Wauck et al., 2017), others collect feedback during usage of the software products for further refinements and continuous improvement (e.g., Oppenlaender, Tiropanis, and Hosio, 2020; Seyff, Ollmann, and Bortenschlager, 2014; Stade et al., 2017). The collected feedback can mainly be split into two groups: qualitative feedback and quantitative feedback (Haug & Maedche, 2021a). While qualitative feedback represents mostly texts or videos, quantitative feedback is collected via votes or ratings. The scope of the feedback also differs between existing systems. Most systems collect feedback on non-functional attributes, such as aesthetics and human values. However, the collection of feedback on content and functional attributes is also supported. Similar to other crowdsourcing systems, crowd-feedback systems also differ in the crowdsourcing configuration, which here comprises the type of crowd (anonymous, users, students, convenience) and the incentive (money, involvement and improvement, interest, and social compensation, credits, and gamification). Crowd-feedback systems differ also in their design characteristics. Haug and Maedche (2021a) thereby identified nine design features: questionnaires, free text field, categories, selection, direct manipulation, collaboration, markers, context, and recording. Finally, it has been shown that crowd-feedback systems do not only have positive effects on the process, but also on outcomes such as feedback quality and quantity, and the resulting design.

Crowd-feedback systems provide multiple benefits for software and user interface (UI) designers to continuously evaluate the software designs during the development process. However, they have downsides as well. Design features might provide the ability to collect design feedback focused on dedicated aspects depending on the situation and enable de-

signers to receive high-quality design feedback. Although the feasibility of crowd feedback for various kinds of designs and systems has been proven, there is still a lack of research on the individual effects of specific design features on feedback quality and quantity, as well as the behavior and engagement of feedback providers. Consequently, it remains unclear how and when to apply these features in crowd-feedback systems.

3.2.3. Crowdworker Perspective in Crowd-Feedback Systems

In summary, one can distinguish two perspectives in crowd-feedback systems: 1) the perspective of feedback requesters that design a system, create crowdworking tasks, and request feedback and 2) the crowdworkers' perspective who conduct the tasks and provide the feedback. Research and practice so far have primarily focused on the development of efficient crowd-feedback systems that generate optimized results for the feedback requester. However, it failed to consider the feedback providers' perspective of the crowdworkers, their experience, and their impact on the feedback outcomes.

Oppenlaender and Hosio (2019) and Robb, Padilla, Methven, et al. (2017) showed that user engagement plays an important role when crowdsourcing feedback. Increasing the engagement of the crowdworkers improves feedback quality and quantity (Oppenlaender & Hosio, 2019; Robb, Padilla, Methven, et al., 2017). A potential explanation is the Theory of Interactive Media Effects (TIME) (Sundar, Jia, et al., 2017). The TIME states that features, sources, and content of software affect the users' perception as well as their behavior. As a core characteristic, according to the TIME, the interactivity of software features impacts user engagement. The interactivity addresses the methods of interactions that are offered (e.g., clicking, scrolling, dragging). As an explanation, the various interaction methods improve the user's mental representation of the software. As a shortcoming, however, higher interactivity also affords greater perceptual bandwidth and might aggravate efficient usage (Sundar, Jia, et al., 2017). The relationship between feature interactivity and the user's absorption in and attitude towards the system is mediated by the ease of use of the software besides its natural- and intuitiveness (Sundar, Jia, et al., 2017). Ease of use is an important factor for the success of crowdworking tasks. Therefore, the application of TIME in the context of crowd-feedback systems might allow us to better focus on the crowdworker perspective (Sundar, Jia, et al., 2017). While increasing the level of interactivity and subsequently the level of user engagement helps to improve the feedback,

better ease of use of the software can be a path towards a higher level of crowdworker experience.

This can also be explained by the concept of information overload (Hiltz & Turoff, 1985). Roetzel (2019) defines information overload as the situation "when decision-makers face a level of information that is greater than their information processing capacity" (p. 480). Being presented with too much information, in our case, multiple options to provide feedback, can lead to people failing to respond to inputs or ignoring information (Hiltz & Turoff, 1985). Consequently, when users are overwhelmed by many options, they might ignore some of them or fail to use them. We believe, that there must be a balance between offering multiple modalities of interaction to increase user engagement and presenting too many options and thereby overloading users.

3.3. Study 1: Design of a Crowd-feedback System based on the Feedback Provider Perspective

The goal of our paper is to design an innovative crowd-feedback system that addresses both, the crowdworkers' and the feedback requesters' perspectives. While increasing the feedback quality and quantity, we aim to provide an enhanced feedback provision experience for crowdworkers. To do so, we conducted a design study, which was already published as a separate poster (Haug & Maedche, 2021b). In this design study, we derived design principles from literature and evaluated users' experiences with the features in qualitative interviews. Based on the results, we designed and developed the crowd-feedback system *Feeasy*.

3.3.1. Method

In the design study, we, first, derived an initial crowd-feedback prototype based on existing design features from the literature. Subsequently, we conducted semi-structured qualitative interviews with exemplary design feedback providers (i.e., crowdworkers) after an interaction with a crowd-feedback system (see section 3.3.1.2). In the following, we present the methodology of this study in more detail.

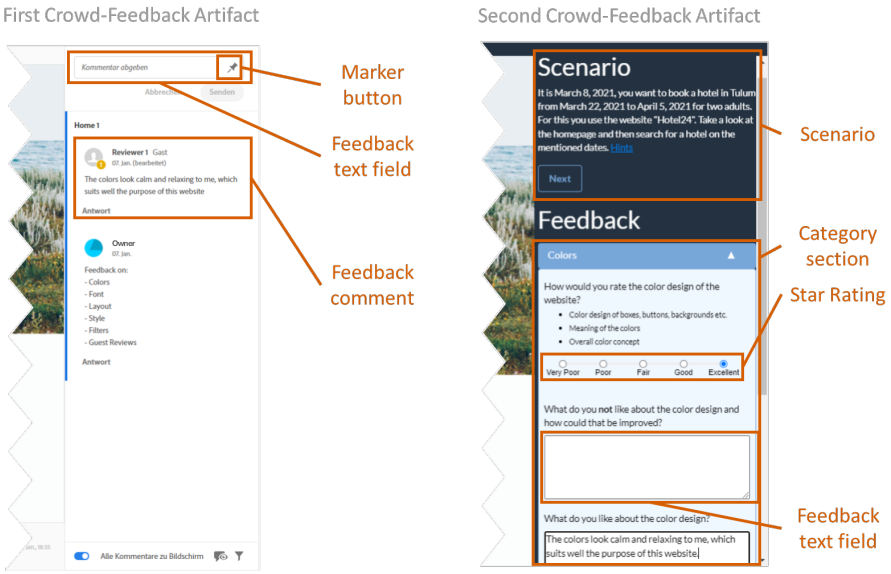


Figure 3.1.: Screenshots of the two feedback panels for the design study. Left: first crowd-feedback artifact (Adobe XD), right: second crowd-feedback artifact (self-developed).

3.3.1.1. Procedure

We recruited ten students for the analysis of crowd-feedback system design features. Four participants were female (six male) and they had an average age of 23.10 years (SD = 2.95). We asked for their level of experience with UI and UX design on a five-point Likert scale. Participants reported little experience on average. For the design study, we derived design features from literature and included them in two distinct crowd-feedback artifacts. We split the participants randomly into two groups of five people. The general procedure for both groups was the same. All participants had to interact with one of the two crowd-feedback artifacts to put themselves into the situation of providing feedback and experiencing the design features. Their task was to provide feedback on a low-fidelity prototype of a hotel-booking website. We decided on this prototype because we assume that previous experiences with hotel booking websites among participants are similar. The prototype consisted of four different subpages and blue boxes showed participants where to click. Participants could interact as long as they preferred. Most participants needed 20 - 30 minutes to complete the instructions and the interaction itself. Afterward, they participated in semi-structured qualitative interviews which took around 20 minutes. The qualitative interviews mainly focused on understanding how participants perceived the design features of the crowd-feedback artifacts they interacted with. However, we also asked interviewees about their opinions on further design features that were not included in one

of the two crowd-feedback artifacts (e.g., collaboration and voice input). For participation in the whole study, we paid everyone $11.3. The interviews were conducted in German and then translated to English.

3.3.1.2. Study Artifacts: Crowd-Feedback Systems

We decided to let participants interact with two different crowd-feedback artifacts to be able to receive opinions on multiple design features. Both crowd-feedback artifacts are shown in Figure 3.1. Design features of crowd-feedback systems can generally be split into nine different types with either the goal to collect feedback (feedback collection mechanisms) or to enrich and improve the feedback (interactivity cues) (Haug & Maedche, 2021a). These nine design features are free text field, questionnaire, categories, selection, direct manipulation, context, markers, recording, and collaboration (Haug & Maedche, 2021a). We describe all features in Table 3.1.

Table 3.1.: Overview of all design features of crowd-feedback systems according to Haug and Maedche (2021a).

	Feature	Definition	Examples
Feedback Collection Mechanisms	Free Text Field	A single text field for feedback without any specific questions	Seyff, Ollmann, and Bortenschlager (2014), Y. W. Wu and Bailey (2021), and Yen, Dow, et al. (2017)
	Questionnaire	A series of questions to answer	Ayalon and Toch (2019), Nebeling et al. (2013), and Oppenlaender, Kuosmanen, et al. (2021)
	Categories	Categories or rubrics to add feedback comments to	Easterday et al. (2017), Schneider et al. (2016), and Yuan et al. (2016)
	Selection	Feedback is provided by selecting items (e.g., rating or voting designs)	Choi et al. (2021), Oppenlaender and Hosio (2019), and Robb, Padilla, Kalkreuter, and Chantler (2015a)
	Direct Manipulation	Design can be edited by feedback providers	Choi et al. (2021) and Oppenlaender, Tiropanis, and Hosio (2020)
Interactivity Cues	Context	Feedback providers receive a context of use (e.g., a scenario or a persona)	Ayalon and Toch (2019) and Y. W. Wu and Bailey (2016)
	Markers	Pins can be put onto the design to indicate which element is meant by the feedback or screenshots/pictures can be added to the comment	Luther, Pavel, et al. (2014), Oppenlaender and Hosio (2019), and Schneider et al. (2016)
	Collaboration	Feedback providers can interact with the feedback of others (e.g., add comments or vote)	Nebeling et al. (2013), Oppenlaender, Tiropanis, and Hosio (2020), and Xu and Bailey (2011)
	Recording	Feedback providers can do voice or video recordings	Dow et al. (2013), Oppenlaender and Hosio (2019), and Seyff, Ollmann, and Bortenschlager (2014)

To reduce the development effort in this exploratory phase, we decided to use an existing

commercial crowd-feedback system in the form of the commenting functionality of the commercialized prototyping software Adobe XD for the first crowd-feedback artifact.

This first crowd-feedback artifact collects the feedback in parallel to the design prototype experience. Thereby, the design prototype is on the left side and a panel to add and organize feedback is on the right side. To add a feedback comment, users can enter their feedback in a text field and submit it. As the feedback is only collected via the text field and no categories or questions are included in the UI to guide the users, the free text field is one respective design feature that is applied in this crowd-feedback artifact (Haug & Maedche, 2021a). After submitting a comment, a new comment box is created. Consequently, all comments are displayed as separate boxes. Thereby, each comment belongs to one subpage of the artifact. Before the study, we added an additional comment that showed users on which aspects feedback shall be provided. Users can also add markers to the prototype to indicate which element their comment is addressing. In general, features that allow feedback providers to annotate the user interface or screenshots by drawing boxes or adding pins (e.g., Luther, Pavel, et al., 2014; Stade et al., 2017; Y. W. Wu and Bailey, 2016; Xu and Bailey, 2014), help not only feedback provides to feel more engaged (Oppenlaender & Hosio, 2019) but especially support developers in understanding the feedback (Seyff, Ollmann, & Bortenschlager, 2014; Stade et al., 2017). The marker feature is according to Haug and Maedche (2021a) the second design feature of this crowd-feedback artifact.

We complemented the design features in the first crowd-feedback artifact with a self-developed second crowd-feedback artifact that contains further design features. In the following, we outline the design of the second crowd-feedback artifact, which is derived from existing literature, in more detail.

General Layout

The general layout is characterized by the parallel arrangement of the design prototype that allows interaction with a prototype on the left side and the feedback panel on the right. This allows for a close direct connection between the prototype experience and the feedback provision and is innovative compared to other recent crowd-feedback systems in practice (e.g., Oppenlaender, Kuosmanen, et al., 2021).

Design Features

The design features which are not covered by the commercial crowd-feedback system that we use in this study, are questionnaire, categories, selection, direct manipulation, recording, collaboration, and context (Haug & Maedche, 2021a). In the following, we want to provide a short overview of the characteristics of each of these features before explaining which design features are implemented in the second crowd-feedback artifact and why we decided on them. Compared to the free text field, questionnaires ask users specific questions about their perceptions of the design prototype to collect feedback. Usually, each question has a text field, where the crowdworkers can enter their answer to this question as their feedback (e.g., Xu and Bailey, 2011). Existing crowd-feedback systems apply categories to structure the feedback, guide the feedback providers, and reduce the analysis time of the feedback for requesters as the feedback is already structured (Schneider et al., 2016; Xu & Bailey, 2014; Yen, Dow, et al., 2017). These categories usually represent different dimensions of aesthetics (Luther, Tolentino, et al., 2015), design principles (Yuan et al., 2016), or impressions of the design (Xu & Bailey, 2014). Categories can be implemented as narrow statements users can select to add a comment (Yuan et al., 2016) or broader topics that tell feedback providers what kind of feedback is required (Schneider et al., 2016). The drawbacks of categories are that they might prevent feedback providers from entering feedback that does not fit into these categories (Easterday et al., 2017; Schneider et al., 2016) or that users might misunderstand the categories and consequently submit wrong feedback. While most studies only use categories as a design element without analyzing their effects, Yuan et al. (2016) focused in their study, especially on how categories affect the way people provide design feedback. They learned that categories enable novices to provide feedback that is nearly as valuable as expert feedback. Additionally, they found that this is caused by categories leading to a better writing style. With the selection feature, we summarize all features that enable feedback providers to select something, e.g., a rating score (Oppenlaender, Kuosmanen, et al., 2021), a statement (Xu & Bailey, 2014) or even a picture (Robb, Padilla, Methven, et al., 2017), to share their feedback. In the educational context, ratings lead to more justifications in the feedback but reduce the feedback quality (Hicks et al., 2016). Collecting feedback via direct manipulation means that users can adapt the UI or at least some aspects of it according to their wishes to tell feedback requesters how they would like to have it designed. Probably due to the high

implementation effort, it is only applied in very few crowd-feedback systems in research (e.g., Oppenlaender, Tiropanis, and Hosio, 2020). The recording feature is usually implemented as video and audio recording of feedback (e.g., Seyff, Ollmann, and Bortenschlager, 2014). In related studies, it was found that overall, written feedback is more comfortable for feedback providers, but audio recordings could be a helpful alternative (Seyff, Ollmann, & Bortenschlager, 2014). Collaboration in the context of crowd feedback usually means that users can react to the feedback of others by voting or rating it (e.g., Nebeling et al., 2013). The last design feature, context, includes all features of crowd-feedback systems that provide the crowd with some sort of context in the form of a narrative or a persona that helps them to better understand the context of use. It has shown that offering crowdworkers context increases their empathy and in turn, improves the feedback quality and quantity (Muñante et al., 2017; Wauck et al., 2017).

We decided to apply categories, selection (in the form of a star rating), and context as design features in the self-developed second crowd-feedback artifact. Combining two feedback collection mechanisms has yet only been done by one other crowd-feedback system (Haug & Maedche, 2021a). These two design features are easy to combine and do not require a complex implementation such as for direct manipulation. We included in our panel seven category sections, one for each category, to enter feedback. Each section contains two text fields, one for positive and one for negative feedback, and a five-point Likert scale to rate the design aspect.

We decided to apply context as a design feature due to two reasons. In this initial design study, we relied on easy-to-apply and agile development which allowed only simple prototypes. The implementation of context is much easier than developing a recording or collaboration feature. Second, it has shown that offering crowdworkers context increases their empathy and in turn, improves the feedback quality and quantity (Muñante et al., 2017; Wauck et al., 2017). However, it has never been analyzed how scenario-based instructions influence feedback compared to simple step-by-step instructions. Therefore, we implemented a scenario that describes users a situation that they should imagine when interacting with the design prototype.

3.3.2. Results

We analyzed the results of the qualitative interviews deductively by categorizing them. We report them along the categories of general experiences, the design features that were included in the two crowd-feedback artifacts, and further ideas for improvement.

General Experience

The participants appreciated the parallel arrangement of the prototype and the feedback panel to provide comments. Participants in the first group valued the intuitiveness of providing feedback in Adobe XD as it reminded them of the commenting functionality in similar commercial tools (*"It [Adobe PDF reader] is similar with the comments if you make any [they are] also on the right side, so to speak"* (T1P4)). In the second group, participants missed being able to add feedback to one specific subpage (*"I thought that there is basically one feedback for each page and not always one for all"* (T2P3)). Therefore, we derived the implication for crowd-feedback system design of intuitiveness in commenting and specificity for logical subpages. Furthermore, we learned that offering crowdworkers to interact with the design prototype and provide feedback in parallel is highly appreciated.

Scenario

Both groups thought that the guidance through the design prototype by a scenario was helpful to them. In the first crowd-feedback artifact the scenario was not included. However, as participants still needed instructions about where to click, we included the scenario in the overall task instructions for the experiment. Consequently, participants in this group had to jump between the browser tab with the instructions and the browser tab with the crowd-feedback artifact back and forth, which they disliked. In the second group, the participants liked that they could always have an eye on their objective and felt the task was more interactive by having the scenario (*"I found the example at the top very helpful, that you don't just click wildly, because not everything is clickable anyway. And so you had a goal in mind that you can just do, just to test it"* (T2P4)). Consequently, scenarios are a helpful design feature of crowd-feedback systems, as long as they are included in the UI of the crowd-feedback system.

Markers

The markers were perceived as highly positive. Participants in the first group were enthusiastic about the markers as they helped them to be more precise and reduce the risk of being misunderstood by the feedback requester (*"I think I'm a bit more concrete [with my feedback], so there's less room for interpretation"* (T1P5)). In turn, participants in the second group missed an option to directly annotate the prototype and the ability to pinpoint specific elements in connection to their comments. In summary, a marker-like feature was highly requested by participants who did not have the marker feature, while participants of the first group appreciated it as it helped them a lot to focus their feedback.

Categories

While in the second crowd-feedback artifact the categories were included as separate feedback sections, the first artifact showed only a list of the categories. The participants perceived the categories as very helpful in both groups. They reduce uncertainties about the relevance of feedback and point out things that one might have missed otherwise (*"Categories [...] ease it for many people to just start and think about it [their feedback]"* (T2P3)). Some participants mentioned that even more specific categories might be better. The participants who used the first crowd-feedback artifact missed a way to show to which category their feedback comment belongs. Therefore, including categories as sections where crowdworkers can add their comments serves feedback providers as guidance and also helps them to organize their feedback comments accordingly.

Star Rating

Participants in the first group missed *"...something simple, which is quick and from which you can get the necessary feedback"* (T1P3) like a rating or voting functionality. Participants in the second group appreciated the effortless feedback and the ability to combine qualitative and quantitative feedback to offer a broader picture (e.g., *"you first assess that [the design] in itself in these five categories and then you can think more about it"* (T2P1)). Consequently, star ratings seem to be valuable to workers as they offer an effortless way to provide additional feedback besides pure feedback comments.

Further Ideas for Improvement

Since our goal was to evaluate design features in simple and quick prototypes, we did not include all design features for crowd-feedback systems that are relevant based on previous research. Therefore, we asked participants about their opinions on design features that were not included in one of the two crowd-feedback artifacts.

Participants in both groups were indecisive about recording audio comments for their feedback. While some appreciate the reduced time and effort (*"...because it's faster and because I can share my thoughts more quickly instead of having to write them"* (T1P3)), others worried about the reduced structure and mentioned that they feel weird when talking in front of the laptop (*"...when writing, you're more likely to rephrase than when I sit down with a voice recorder and record things"* (T2P4)). Based on this feedback, we also asked them about a speech-to-text feature. While some worried about the accuracy of speech-to-text features, others thought it might be a better solution than a pure recording feature. Consequently, although participants were indecisive about whether they would see an overall advantage in using voice input features, especially the speech-to-text feature was appreciated by at least some participants and will therefore be considered in the next iteration of our self-developed crowd-feedback artifact.

Regarding collaboration, most participants agreed that they would get biased when they saw what others wrote. They also thought they would feel insecure about sharing unique feedback or think their feedback was useless when others already reported the same ideas. Interviewee T2P5 stated: *"So when you see what others write, you're immediately biased by it. And obviously, it makes it a little bit easier to write your own feedback, but that's not the information that you want to have and that's our job to give you our own feedback."* On the other hand, some participants said that seeing the feedback of others could inspire them to see the design from a different angle. Overall, we think the identified disadvantages of collaboration combined with the higher effort for feedback requesters to handle the collaboration of multiple crowdworkers outweigh their additional inspiration. Consequently, we will not include this feature in further iterations of our crowd-feedback artifact.

From the design study, we know how users perceive selected design characteristics of crowd-feedback systems. Based on these results, we distilled the relevant features and applied

the results to design the crowd-feedback system *Feeasy*.

3.3.3. Feeasy

Based on the insights of the design study, we iterated our initial self-developed crowd-feedback artifact and developed the crowd-feedback system *Feeasy*. All features of *Feeasy* as well as the expected benefits for crowdworkers and feedback requesters are summarized in Table 3.2. *Feeasy* is designed to improve both feedback quality and quantity on design prototypes but also aims to improve the crowdworker perspective by increasing interactivity, user engagement, and ease of use for the crowdworkers. In the following, we explain the general layout as well as the individual design features of *Feeasy* in more detail.

General Layout

Figure 3.2 shows the final user interface of *Feeasy* which consists of an interactive design prototype on the left side and a feedback panel on the right side. This layout has been appreciated by the participants of our design study as it allows them to interact with the prototype and provide feedback in parallel. This design feature has the main goal to reduce the effort for crowdworkers which in turn might lead to more feedback. We decided to offer only one text field for users to create new feedback comments. New comments are then added to the panel as separate boxes and belong to the subpage of the prototype on which the crowdworker reported the comment. Each box contains a label that indicates the respective subpage. This shall help crowdworkers to organize their feedback and in turn, make it better understandable for feedback requesters. In the following, we present the five key design features that we want to evaluate in the following studies.

Scenario

In our initial crowd-feedback system, we offered a scenario that told users where to click while providing them with a realistic usage scenario. Participants in the design study saw no disadvantages in having the scenario. Offering feedback providers some sort of context increases their empathy and, in turn, improves the feedback quality and quantity of comments (Muñante et al., 2017; Wauck et al., 2017). Therefore, we kept this feature for *Feeasy*. We decided to move the scenario to a separate tab in the panel to keep the layout simple and clean.

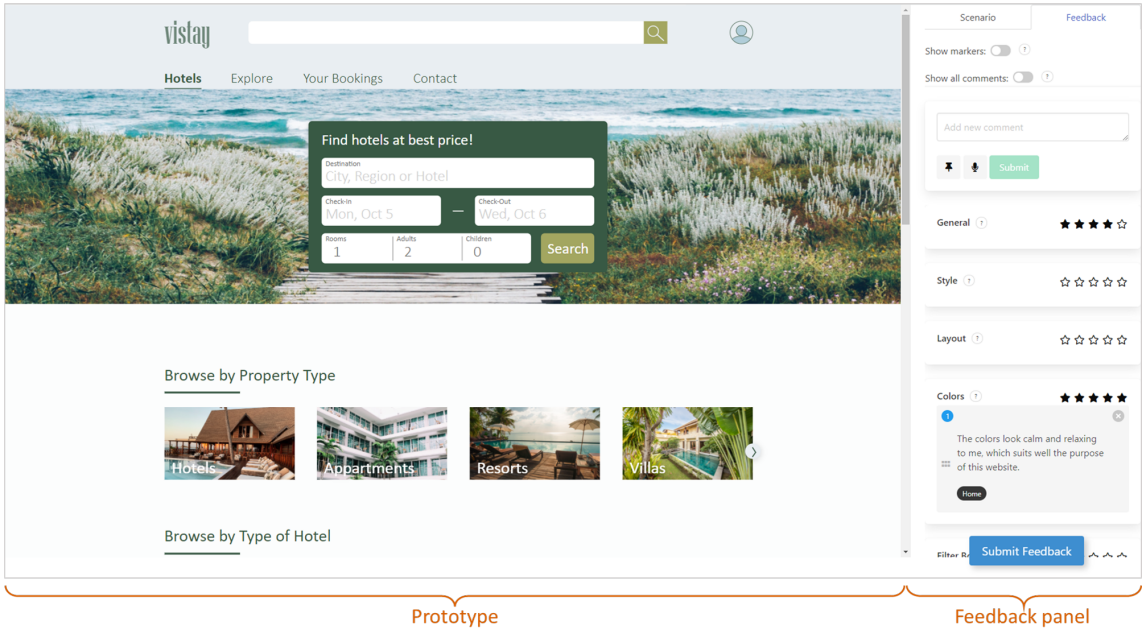


Figure 3.2.: User interface of the interactive crowd-feedback system *Feeasy*.

Table 3.2.: Overview of all features of *Feeasy* and their potential benefits for feedback providers and requesters.

Feature	Description	Provider Benefits	Requester Benefits
General Layout	Design prototype and feedback panel next to each other	Reduced effort	More feedback
	New separate box for each comment	Providers can organize their comments	Feedback that is already split in separate ideas
Scenario	Textual description of an artificial use case for the prototype	Providers know where to click and get more empathetic	Better and longer feedback
Speech-to-text	Speech-to-text input option in text field	Reduce effort and time for providers while still enabling them to edit and structure their thoughts	Longer feedback with more explanations
Markers	Circles with numbers that can be added to the UI and match the number of one feedback comment	Providers can be more specific with their feedback and avoid misunderstandings	More specific and better understandable feedback
Categories	Sections with headlines in which feedback comments can be added via drag-and-drop	Providers can organize their comments and focus on the aspects on which feedback is required	More relevant and focused feedback and comments already organized in categories
Star Rating	Star rating for each category	Providers have a quick and easy way to share additional feedback	Additional quantifiable feedback and more justifications

Speech-to-text

Participants in the design study were mainly indecisive about using a voice input feature. In related studies, it was found that overall text is more comfortable for feedback providers,

but audio recording could be a helpful alternative (Seyff, Ollmann, & Bortenschlager, 2014). Consequently, we decided to offer a speech-to-text feature as an optional input mechanism for feedback. The speech-to-text feature enables users to dictate their feedback. When they click on the microphone button *Feeasy* starts to listen and directly transfers the speech into text. Users can then still edit the text in the text field.

Markers

As markers were found to be helpful for crowdworkers to be more specific and avoid misunderstandings, we implemented them in *Feeasy*. As already explained, markers help not only feedback providers to feel more engaged (Oppenlaender & Hosio, 2019) but also support developers in understanding the feedback (Seyff, Ollmann, & Bortenschlager, 2014; Stade et al., 2017). In our case, the user interface can be annotated with small circles with numbers that belong to one comment box. With the circle, users can indicate which element of the user interface the respective comment is addressing.

Categories

Categories in which users can add respective feedback comments not only enable users to organize their thoughts but also provide value to designers as the collected feedback is already split into categories. It has also shown, that categories help novices to provide better feedback (Yuan et al., 2016). As participants in our design study liked being able to address specific categories and organize their feedback, we kept this feature for *Feeasy* and just adapted it to the improved layout. We included the categories in *Feeasy* as separate sections in which comment boxes can be added via drag-and-drop. We decided on categories that mainly focus on aesthetics (layout, color, font, style), and one category that addresses a specific design element (filter bar).

Star Rating

Participants seemed to appreciate the quick and easy way to share feedback with a quantitative evaluation. Additionally, feedback requesters profit from having additional quantifiable feedback that summarizes the qualitative comments. In *Feeasy*, the quantitative evaluation is included as star ratings. Each star rating is attached to a category. Users can then rate how well they assess each category on a scale from one to five.

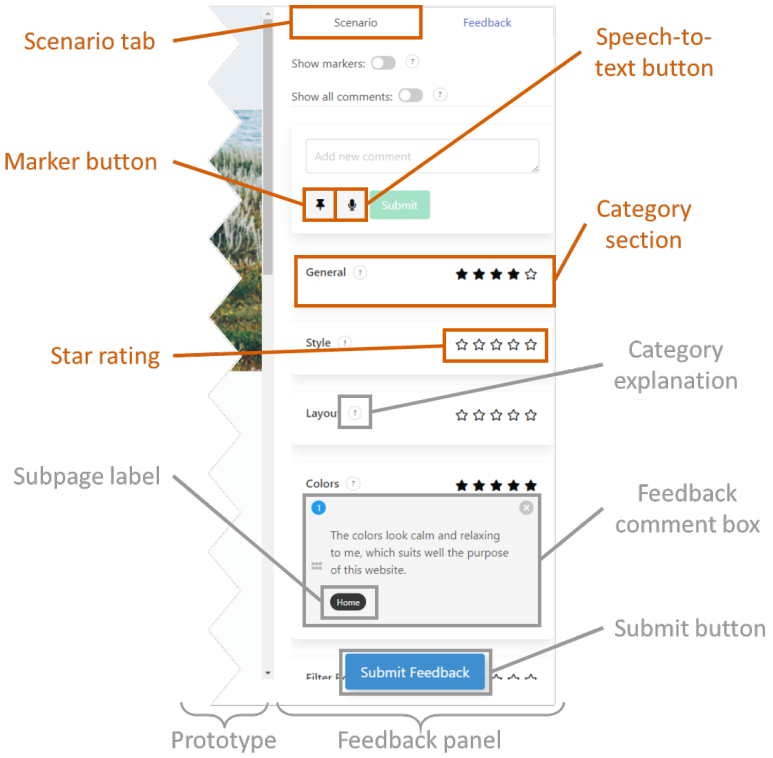


Figure 3.3.: Feedback panel with an explanation of the general layout (gray) and our five key design features (red).

3.4. Study 2: Evaluation of Individual Design Features of Feeasy

In our first study, we started by collecting insights on the distinct effects of innovative design features for crowd feedback from the feedback providers' perspective. Based on the results, we designed the crowd-feedback system *Feeasy*. The primary goal of the second study was to investigate how each individual feature of *Feeasy* impacts the feedback quality and quantity as well as crowdworker perceptions. Specifically, we compared the individual features with a basic version (no features) and a full version (all features) of *Feeasy*.

3.4.1. Method

To evaluate the individual design features of *Feeasy* on the crowdworker perceptions in terms of perceived interactivity, user engagement, and ease of use as well as the feedback quality and quantity, we collected design feedback on a fictitious hotel booking website prototype, through a human-intelligence task (HIT) on the crowdworking platform Prolific. Since our goal is to evaluate the effect of each of the five design features individually and, additionally, to compare the results with a baseline and a full version of *Feeasy*, we derived seven treatment conditions in this study.

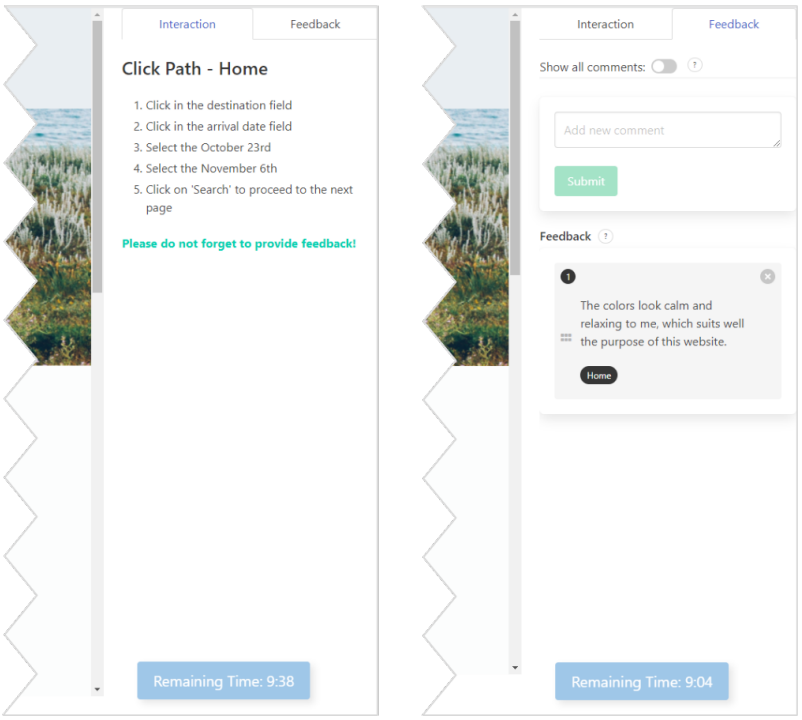


Figure 3.4.: Basic version of *Feeasy* without the five key design features.

3.4.1.1. Procedure

We implemented seven instantiations of *Feeasy*: (1) Full (F), (2) Basic (B), (3) Scenario (S), (4) Speech-to-text (R), (5) Markers (M), (6) Categories (C), and (7) Star Ratings (Q). The basic version is displayed in Figure 3.4. For all five treatments with a single feature, the treatment instantiations looked like the basic version plus the respective feature as implemented in the full version (cf. Figure 3.3). For example, for the speech-to-text treatment, the *Feeasy* interface looked like the basic interface (cf. Figure 3.4) with just the microphone button added below the text field for adding comments. Only for the star rating, we had to additionally include the categories as the rating is always attached to a category. All variants that did not include the scenario contained an interaction tab instead that showed a step-by-step list for each subpage of the prototype to tell users where to click. When starting the HIT, participants received an introduction to the specific version of *Feeasy* according to the respective treatment as well as a short training on how to provide high-quality design feedback by addressing relevant feedback aspects. After the introduction, each participant was randomly assigned to one of the seven treatment conditions and experienced the treatment phase. During the treatment phase, the participants had to walk through a basic design prototype. Similar to study 1, this was a fictitious

hotel booking website (cf. Figure 3.2). The prototype consisted of four subpages on which the participants gave feedback. To conduct the task the participants had to use *Feeasy* and provide design feedback comments in the treatment phase for at least ten minutes. After ten minutes they were allowed to submit their design feedback and move on to the next step. The collected design feedback comments, as well as further information (e.g., for which comments the speech-to-text feature was used), were stored in a database. After the treatment phase, participants answered a quantitative questionnaire that asked for their perceived user engagement, perceived interactivity, and perceived ease of use of the experimental prototype *Feeasy*. Afterward, to additionally collect qualitative data, the participants were offered to book an appointment to participate in an interview. Finally, participants received a debriefing and their compensation.

3.4.1.2. Participants

For study 2, we recruited 210 participants via Prolific. Of the participants, 48.10% were female (51.43% male) and the average age was 25.70 (SD = 7.62), while the youngest participant was 18 and the oldest 65. On average, the participants reported limited experience in UI/UX design on a seven-point Likert scale (M = 2.30; SD = 1.43). Since the task was to provide design feedback on a hotel booking website prototype, we asked for their frequency of visiting hotel booking websites on a seven-point Likert scale. Their experience with this was limited as well (M = 2.93; SD = 1.24). The participants were distributed on the seven treatment conditions with 29 to 31 participants per treatment. For the task, participants received compensation of $5.0. On top of that, we provided flexible compensation to create a realistic crowdworking scenario and to motivate crowdworkers. The flexible payment was a $1.0 bonus given to participants that ranked within the 30.0% best participants in terms of quality and quantity. Eventually, we paid the bonus to everyone who faithfully completed our task. This resulted in a payment of $6.0 for around 30 minutes of work which is above the German minimum wage ($11.0 per hour). 28 participants took part in the subsequent qualitative interviews, at least three per treatment. The interviews took between 15 and 20 minutes and participants were compensated with an additional payment of $5.0. We removed two of the interviews from the following analysis due to low quality and misunderstandings caused by the language barrier of the crowdworkers.

3.4.1.3. Data Collection & Analysis

We collected data in two ways. First, we collected quantitative data via the questionnaire for three constructs: perceived interactivity (consisting of fifteen items by Liu (2003)), perceived user engagement (consisting of seven items by Webster and Ho (1997)), and perceived ease of use (consisting out of six items by Davis (1989)). For perceived interactivity, we removed the sub-construct synchronicity since all treatments of *Feeasy* should perform similarly. We also removed all items related to feedback or communication with the website since communication with *Feeasy* was not relevant to this study. This led us to a final set of seven items for perceived interactivity.

Second, we analyzed the feedback comments collected from the crowdworking task on their feedback quality. Before the analysis, we excluded feedback comments from participants who failed one of our three attention checks (i.e., in the form of attention questions in the questionnaire: *"If you are carefully filling out the survey, please select strongly disagree."*). Further, we removed participants who wrote no feedback comments. In the full treatment, we asked participants to rank the five features according to their importance for the feedback. To analyze the feedback comment quality we created another HIT in which UI-design-skilled crowdworkers assessed the quality of the design feedback comments. For this HIT, we again used the crowdworking platform Prolific since it allows us to filter for workers with UI design skills. We recruited 160 workers with UI design experience ($M = 4.35$, $SD = 1.61$, based on a 7-point Likert scale). Since the assessment of feedback quality required prior knowledge about the prototype and relevant design feedback dimensions, the participants initially received an overview of *Feeasy* and the specific aspects they should consider in their assessment.

Subsequently, each feedback comment provided in the initial HIT was analyzed by three participants on the quality categories of helpfulness, specificity, relevance, sentiment, and objectivity. Complementary to the text comment, participants received additional information about potential markers that were added and to which category and subpage the comment belonged. Following previous work on the assessment of feedback quality (e.g., Xu and Bailey, 2014; Yuan et al., 2016) helpfulness serves as a measure for the overall quality, while the remaining four constructs represent detailed constructs to assess design feedback (Krause et al., 2017; Oppenlaender, Kuosmanen, et al., 2021). A description of each quality construct can be found in Table 3.3. The feedback quality value for each con-

struct was assessed by taking the average from the distinct ratings of the three individual crowdworkers.

Table 3.3.: Explanation of feedback aspects.

Feedback Aspect	Description
Helpfulness	Helpfulness addresses the overall quality of the feedback comment.
Sentiment	Sentiment assesses if the comment is rather addressing a problem of the design (lower rate) or if it is praising the design (higher rate). Simple statements without judgment should thereby be neutral.
Objectivity	Objectivity evaluates how much the comment is based on facts and not only personal beliefs, opinions, and preferences.
Relevance	Relevance assesses how relevant the comment is to further improve the design of the hotel booking website. Thereby, crowdworkers should consider the categories on which we collected feedback and the limitations of the prototype (e.g., functionalities).
Specificity	Specificity addresses how specifically the feedback has been phrased. This includes how clearly it describes the element it is addressing and its positive or negative aspects.

For the qualitative analysis, we conducted semi-structured qualitative interviews with the participants who were willing to provide their insights after the HIT. The questions in the qualitative interviews focused on crowdworkers' experiences with their version of *Feeasy* in general and each feature in particular. Additionally, we asked participants about their procedure to provide feedback and ideas for further improvement. We analyzed the feedback through a deductive thematic analysis following Braun and Clarke (2006) based on the TIME theory. To facilitate the analysis we organized the results around three categories of general experiences of *Feeasy*, its positive aspects, and its negative aspects regarding the categories of the TIME theory (i.e., interactivity, engagement, ease of use, feedback quality, feedback quantity).

3.4.2. Results

3.4.2.1. Quantitative Analysis

To assess the participants' perceptions, we analyzed the responses to questionnaire items. To assure the internal consistency of latent constructs, we assessed outer factor loadings and Cronbach's alpha with a cutoff at 0.7 and 0.6 (Hair et al., 2014; van Griethuijsen et al., 2015). Since not all constructs did meet these requirements we removed perceived interactivity items five and six having Cronbach's alpha then ranged from 0.68 to 0.78. Afterward, scales were averaged. To assess the effect of the experimental treatment conditions (basic vs. full treatment), we conducted a multivariate analysis of variance (MANOVA) with the

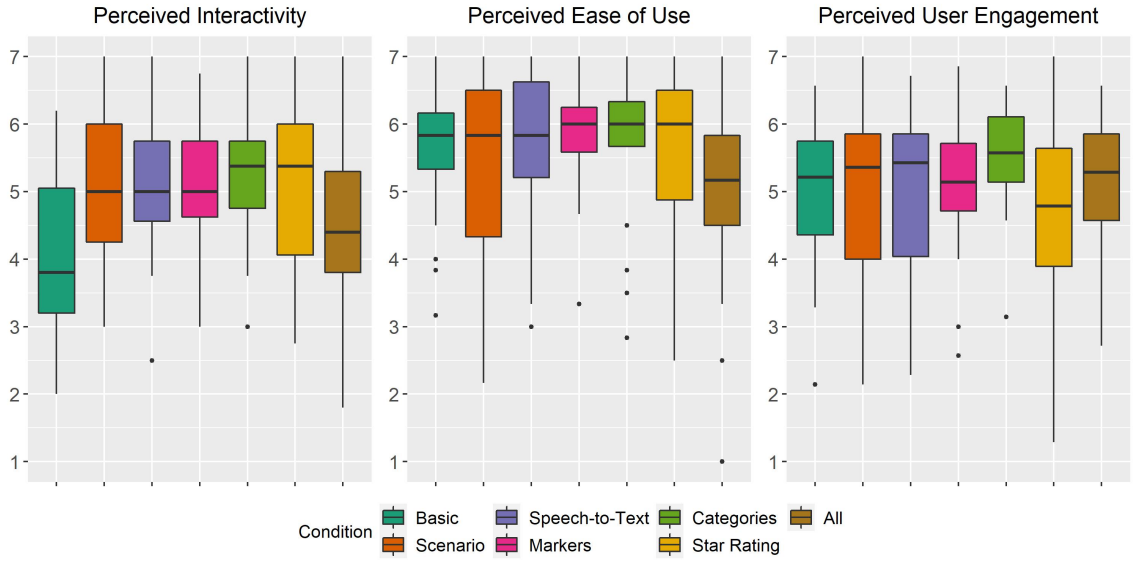


Figure 3.5.: Boxplots of perceptions of interactivity, user engagement, and ease of use measures of the crowdworkers.

three perceptive measures and the feedback quality and quantity assessments as dependent variables. Since the variables under investigation violated the assumption of univariate and multivariate normality, we conducted a nonparametric rank-based MANOVA using the R software package rankMANOVA (v. 0.0.7) (Dobler et al., 2020). The results of the rank-based MANOVA for analyzing nonparametric data did not reveal a significant effect of the treatment conditions on the dependent variables. Furthermore, we conducted an aligned rank transform (ART) for nonparametric factorial analyses of variance procedures using the R package ARTool (v. 0.11.1) (Wobbrock et al., 2011). Results show no significant results besides a significant effect of ease of use ($p < 0.05$) between the full and the basic version without design features and a significant main effect for specificity ($p < 0.05$). To complement the quantitative analysis, we, then, pursued a thorough descriptive analysis of the data.

Figure 3.5 shows that the perceived interactivity for all five treatments with only one feature (except the star rating treatment, which includes two features) is higher than for the basic treatment and the full treatment. Thereby, the perceived interactivity in the scenario and speech-to-text treatment is still lower than for the marker, category, and star rating treatments. The perceived ease of use is in all individual feature treatments similar to the perceived ease of use of the basic treatment and higher than for the full treatment. The results for perceived user engagement differ between the five treatments. While the

perceived user engagement for speech-to-text, categories, and the scenario is higher than for the basic and full treatments, the perceived user engagement for markers and star ratings is lower. The highest perceived user engagement was achieved for categories, while the lowest was the star rating treatment, which also included the categories.

The overall quality which was described by the helpfulness of the feedback comment is stable across the individual feature treatments and lower for the basic and especially the full treatment condition. Regarding the sentiment, categories have led to more positive comments compared to the other treatments. However, the differences between the treatments were only marginal. Regarding objectivity, there was no difference between the treatments. All features and combinations of features have led to medium objective feedback comments. The relevance again was the lowest for the basic and full treatment, while there is no difference between the other five treatments. Finally, the specificity is the highest for markers and the lowest for the full treatment. The number of comments per crowdworker was the highest in the category treatment. The lowest number of comments was achieved for the scenario and the full treatment. Regarding the comment length, the results of all treatments were similar with comments having between 70 and 170 characters. Only the number of characters in the full treatment was lower than the rest.

To analyze the results of the ranking task of the full treatment, we calculated Kendall's W to know how much the participants agreed on their ranking. The Kendall-W-Test is a non-parametric statistical test that compares the distributions of three or more related variables and analyzes if these variables are significantly different from one another. Kendall's W can range between 0 (no agreement) and 1 (full agreement). For the test, we transformed the ranking into ordinal values from one to five with one meaning the feature was ranked the most important and five meaning the feature was ranked the least important. We received a Kendall's W of 0.31 which indicated a rather low agreement among the participants. Figure 3.6 presents stacked bar plots of the rankings of the five design features. The scenario feature was on average ranked the most important ($M = 2.06$) and the recording feature the least important ($M = 4.45$). The star rating and the markers were perceived as similarly important and the categories as slightly less important.

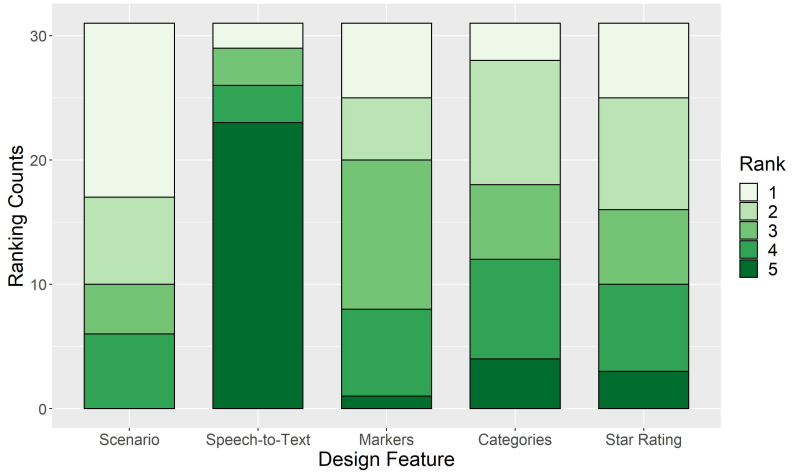


Figure 3.6.: Stacked bar plots of ranking of the five design feedback features (from rank number 1 (best) to rank number 5 (worst)).

3.4.2.2. Semi-structured Interviews

In this section, we summarize the insights that we gained during the 26 qualitative interviews that we conducted with crowdworkers who successfully completed the feedback task. We first report their overall experiences and how they proceeded to provide the feedback, and then we describe their perceptions of the five individual design features of *Feeasy*.

Overall Experience

Overall, the participants enjoyed the interaction with *Feeasy* irrespective of the treatment condition they experienced. All participants in the basic treatment condition appreciated that *Feeasy* was easy and straightforward to use (*"It's not overloaded with anything, which is great. It's really awesome"* (B2)). One participant even stated that the ease of providing feedback made her "*very willing to give out [...] as much feedback as I could because it wasn't frustrating*" (B3). On the other side, participants in the full treatment condition mentioned the need for more time to get familiar with the system and the options: "*I thought I could maybe have spent a little bit more time looking to give feedback if I've spent less time trying to work out how to work the panel*" (F1). One participant even got frustrated about the interface because s/he didn't understand how to interact with it and stated that s/he would have liked to have a practice before the task to feel more comfortable. In the full treatment condition, participants reported multiple times that they missed using one of the five design features accidentally although they remembered being introduced to the features. For example, F1 mentioned, *"It wasn't immediately*

apparent in the panel that that [the speech-to-text] was an option. It was just a small icon from what I remember".

The remaining aspects that crowdworkers favored or disliked varied a lot between the individual participants and therefore seem not to be related to a specific treatment. Three participants mentioned that they liked the ability to see all previous comments in the feedback panel at once so that they were *"able to keep track of all the comments I've made in the previous time"* (S4). Further participants appreciated, especially in the categories treatment, that they were able to edit their comments after submitting them as they *"...kept finding different stuff that I wanted to add"* (C2). Crowdworkers enjoyed that the system was similar to other systems they use for work and reported that they perceived *Feeasy* to be interactive and felt engaged by it. A lot of criticism was around the interactivity of the design prototype itself (i.e., the hotel booking website). Crowdworkers stated that being able to click on more things would have led to more feedback (*"If we weren't limited to the testing if we weren't limited to features, I think that would have improved our feedback results"* (R2)).

The second main point of criticism was about the parallel layout. One interviewee recommended placing the feedback panel somewhere else, as *"...it's kind of like narrow and I couldn't see everything clearly"* (Q2). The interviews showed that crowdworkers followed different approaches to identify design issues and report feedback. Very common was that they put themselves in the shoes of another person (e.g., *"...a generic person"* (F3), *"...their grandma"* (B2)) or reported everything that seemed counter-intuitive to them or did not meet their expectations towards the design prototype. In detail, they often looked *"...for things that were different and similar to websites that I know"* (F3) because *"...if we don't have what to compare, I don't think we can choose what is best, what is worse, what can improve"* (C1). Crowdworkers in treatments that did not include the category section still used the categories that were provided in the instructions to make sure that they addressed every category. Participants without the category feature *"...just reported everything that came to their mind"* (B3), while participants with categories used them to decide which feedback is relevant for the feedback requester.

Some workers reported that they experienced problems with identifying issues with the prototype's UI or *"...find words to explain what is going wrong on the page"* (M2). To sum up, crowdworkers were very positive about their experience with *Feeasy*. Overall, they

appreciated its simple and intuitive UI (*"I feel like that your feedback box is perfect for every user because it's simple and straightforward"* (R2)).

Table 3.4.: Summary of crowdworkers' perspectives on the design features derived from the interviews.

	Advantages	Disadvantages
General Layout	All submitted comments visible Comments can be edited Similar to other tools Interesting, interactive, and engaging	Only one scenario/click path included Lack of guidance
Scenario	Goal-oriented Better focus Equalizes previous knowledge Interaction more real Clear and straightforward	Feedback focused on click path Lower readability Hard to understand
Speech-to-text	Inclusive Higher quality (more comprehensive) Less time and effort More feedback	Option not clear Slower than typing Less organized Feedback more casual Feedback less reliable Not always convenient
Marker	Easier Comments more specific and detailed Small items highlighted	Redundant Feedback too specific
Categories	Better organization Inspiration and guidance Feedback more complete Better focus	"General" too general Moving comments is annoying Less generic comments No methodological guidance
Star Rating	Flexible and easy Good summary of comments Relativizes harsh feedback	Additional effort

In the following, we present detailed insights for each of the five design features.

Scenario

Crowdworkers appreciated that the scenario feature was clear and straightforward and provided them with a goal to focus on. However, the interviews showed that crowdworkers did not perceive the scenario as a design feature. Crowdworkers felt that the scenario made their feedback more real and relevant to the designer (*"With that text, we can give better feedback because we imagine ourselves like these people like we are going to travel"* (S3)). Additionally, they liked that they knew on which parts of the user interface they should focus (*"Maybe it somehow points my attention to specific things. That might have been helpful"* (S5)). One interviewee even stated that the scenario might be especially helpful *"...for not so experienced travelers or new travelers"* (S3). However, F1 reported problems with the scenario instructions since s/he did not find all subpages of the design

prototype and finally gave up. Some participants also would have preferred bullet points instead of a block of text to make it more readable. Participants with the step-by-step instructions did not report any problems, however, they mentioned the creative limitation of the restriction to one user flow (*"I realized that even if I was tempted to play around a bit with the website, I needed to focus on the goals. So, my focus was on actually completing the steps even though [...] you just automatically want to just hover over the little things and see what is what."* (B1)).

Speech-to-text

Our log data showed that none of the participants actively used the speech-to-text feature. Nevertheless, due to its sole presence, most interviewees saw its advantages as *"...your spoken word is better than written text"* (F1). Therefore, we asked crowdworkers about their reasons for not using this feature and what advantages they still see in entering feedback via speech-to-text. Participants provided various reasons why they did not use it: they were not drawn to it (F1), they did not use it because they did not want to disturb the people around them (F3), or they expected the speech-to-text feature to malfunction because of their accent or the quality of their voice and feared to have to recheck all the feedback as the speech-to-text feature might misunderstand them (*"I'm sure it wouldn't catch anything"* (R2)). Furthermore, crowdworkers expected their feedback to be less organized and more casual when using the speech-to-text feature (*"I feel like maybe when I type, I'm more formal in my phrasing than if I was speaking"* (R1)). Interviewee R3 did not use the speech-to-text feature because s/he assumed that s/he types much faster than s/he speaks. Interestingly, the other two interviewees of the speech-to-text treatment reported the major advantage of the speech-to-text feature in the reduced time and effort for providing feedback as it corrects the spelling and feels for them to be easier than typing (*"...overall, I can say it's much easier to use than typing"* (R2)). Additionally, it makes the feedback provision process more inclusive as also crowdworkers that have problems with fast typing, for example, caused by a disability, could easily provide feedback. Furthermore, the speech-to-text feature could lead to time savings, as it corrects the spelling and feels for some crowdworkers to be faster than typing. Interviewees also stated that they could imagine that *"...there will be more explanation when I say it vocally than typing"* (R2) and that they *"...probably would have given more feedback"* (F1).

Marker

The prevailing perception of the markers was positive in the crowdworkers' interviews. Participants mentioned occasionally that the use of markers is in some cases redundant (*"If I'm describing icons or the selection menu I think it doesn't require pinpointing with a marker"* (M1)), and that they might sound too focused when using them (*"I didn't want to sound like to focus one particular thing."* (M1)). On the other hand, the comments of the participants got more specific and detailed and the markers made it possible to highlight very small items like icons: *"It allows you to pinpoint the specific areas which creates greater visibility and you know there are more layers on your feedback"* (M3). Interestingly, participants without the markers stated multiple times that they would like to *"...just click on something and it being a reference to my feedback"* (B3) and thereby indirectly mentioned the benefit of the design feature of markers.

Categories

The crowdworkers perceived the categories as beneficial for the feedback provision. They liked that the categories helped them to better organize their feedback (*"I think it was just more concrete and more structured than it would be without it"* (C3)) and used it for inspiration and guidance (*"I don't have to wonder what should more I write. [...] I have something to each topic [...] and then I just kept adding if I found something"* (C3)). They felt like the categories helped them to provide more complete and specific feedback and focus on the important aspects of the evaluation process (*"Knowing it [the filter box] was a focus and the main topic of a category, I was able to spend more time on that and it definitely helped for sure"* (C2)). F3 would have liked to have even more categories to provide feedback to. On the other hand, participants stated that *"...it was quite hard to move comments into the specific subsections"* (F3) which annoyed them. One crowdworker also had problems with categories as s/he was not sure what kind of feedback was expected from them (*"I don't think they helped me very much in how to analyze"* (C1)). Finally, using categories too extensively might lead to less feedback that addresses general aspects like the overall style of the website (*"If you put too many categories then you risk of focusing too much on these specific things and not focus on the general website and not give complete feedback and comments on how the website looks as a whole"* (C2)). At the same time, interviewees did not like the *General* category as *"...the 'general' category gets*

very general. And what does that mean? It's not really very specific" (F1)).

Star Rating

In the star rating treatment, crowdworkers could use the categories for their feedback and add an additional star rating for each of the categories. The only disadvantage that crowdworkers reported about the star rating was that it was an additional effort compared to providing just a text comment. However, they still stated that they *"...didn't really find it necessary [...] but it was nice because I could sum up what I thought about it"* (Q2). F1 tried to make sure that *"the star rating was compatible and mirrored the feedback that I had given"*. Furthermore, crowdworkers liked that the star rating provided them the flexibility to rate each category differently. They perceived the star rating to be very easy and good for providing a summary of the text feedback (*"I think it's a very quick way to just say 'OK, this is what my general thoughts were'"* (F3)) and relativizing feedback that might sound too blunt as *"...the stars are very international"* (F2). Further, they thought a good star rating might soften very critical feedback so that the requester understands that despite the criticism the feature is good (*"If I will not do ratings, then nobody will be able to understand how much I really like it and how much I did not like it"* (Q1)).

3.5. Discussion

In this section, we synthesize and discuss the insights that we gained in our two studies. Thereby, we put the feedback provider (e.g., crowdworker)) perspective in the focus for crowd-feedback system design and highlight the interplay with the feedback requester objectives. Our results provide evidence that more design features are not significantly better than applying no design features at all. Furthermore, we learned that especially scenarios were appreciated by crowdworkers, and single-feature treatments performed better than the full and basic treatments in terms of crowdworkers' perceptions, feedback quality, and feedback quantity. Based on these insights we derived design implications for the design of crowd-feedback systems that align both the feedback requester objectives and the crowdworker experience.

3.5.1. Number of Design Features

According to the TIME theory by Sundar, Jia, et al. (2017) the perceived interactivity of *Feeasy* should increase when combining multiple features as this offers feedback providers

more interaction opportunities. However, in our case, the perceived interactivity increased when applying one feature compared to no features, but it decreased for the full treatment with five features. It also showed that the perceived ease of use is stable across the basic treatment and the treatments for the individual features, but lower for the full treatment. However, as the perceived user engagement is similar across all treatments this could support the statement of Sundar, Jia, et al. (2017), that additional factors such as naturalness, intuitiveness, and ease of use are important mediators for the relationship between perceived interactivity and perceived user engagement.

We also assume the perceived interactivity to have an impact on the resulting feedback quality as the feedback from the basic and full treatments which achieved the lowest ratings for perceived interactivity also performed the worst in terms of feedback quality. This holds in particular with the helpfulness category. We found evidence in our interviews that crowdworkers were overwhelmed by having so many options which can be explained by the concept of information overload (Hiltz & Turoff, 1985; Roetzel, 2019). Crowdworkers, therefore, needed some time to get familiar with them. This might have harmed the feedback that these crowdworkers provided. However, we assume the relationship between the three perception constructs and the resulting feedback quality and quantity to be more complex than we expected initially. Based on the insights from the interviews, we believe that additional aspects such as learnability and understanding need to be considered when designing crowd-feedback systems. In the crowdworking context, the simplicity and clarity of tasks and instructions are key.

While the feedback quality for the full treatment is lower than the feedback quality of the single-feature treatments, the full treatment still provides multiple additional benefits to feedback requesters. Most of the feedback is already categorized, some comments include markers that might increase the comprehensibility and in addition, a quantitative assessment is provided. The feedback quantity is slightly lower for the full treatment than for the other variants. This could be caused by the additional effort and time workers had to spend on learning multiple features. As workers spend more time learning the features they have less time to spend on writing feedback comments. Consequently, applying multiple features might have a negative impact on crowdworkers' perceptions, as well as the feedback quality and quantity.

The results of our studies show that a well-dosed application of certain design features has

beneficial effects on crowdworkers and their feedback. As the combination of features might decrease the perceived ease of use and therefore negatively impact overall crowdworkers' perceptions, our recommendation is to use only the design features that are necessary to fulfill the requirements of the design evaluation.

3.5.2. Individual Features

The main goal of study 2 was to compare the effects of the five design features (scenario, speech-to-text, marker, categories, star rating) of our crowd-feedback system *Feeasy* on crowdworkers' perceptions as well as feedback quality and quantity. Overall, the perceived user engagement of crowdworkers and the helpfulness of the resulting feedback comments did not seem to be directly related to each other. For example, the perceived user engagement in the basic treatment was similar to the perceived user engagement in the marker treatment. However, the helpfulness of the feedback comments in the marker treatment is rated much higher than that of the feedback comments in the basic treatment. The same applies to the relationship between perceived user engagement and feedback quantity. While the perceived user engagement is the lowest for the star rating, the feedback quantity is similar to the other treatments.

Considering this, the independence between feedback quality and quantity and the crowdworkers' engagement in our study may imply that in the crowd-feedback context, other factors play a crucial role in increasing the feedback quantity and quality. We learned in the interviews, that crowdworkers appreciate clear and easy features as well as structured guidance in performing their tasks. Additionally, we understood that some workers were insecure about the requesters' expectations of their feedback which might have negatively influenced their feedback quality and quantity. Consequently, additional influencing factors on the feedback quality and quantity might be how well users understand their task and how well the system supports them in expressing themselves and guides them through the feedback task. From the feedback requester's perspective, the objective is to receive feedback with high quality in large quantities. Looking at our quantitative results for the feedback quality and quantity, we were not able to identify significant differences between the treatments. Therefore, we will connect the descriptive results with the interview insights in order to understand the effects of the individual features. Participants of the full treatment ranked the scenario feature as the most important for providing feedback. We

assume the reason was that they did not understand how to interact with the prototype without knowing where to click. Consequently, they perceived the scenario as essential to provide feedback. It also might have helped to make the feedback situation seem more natural. When leveraging the scenario, feedback requesters must consider that the scenario leads to workers' feedback being more focused on specific elements and features. Consequently, the scenario is helpful in particular, when feedback is required for a specific part of the design prototype like a new feature. Looking at the crowdworkers' perceptions, the categories feature performed the best for all three perception constructs (perceived interactivity, perceived ease of use, and perceived user engagement). Remarkably, the categories were the only feature in which crowdworkers' reported usability issues in the interviews. Crowdworkers reported that they had problems moving the comment boxes into the right category sections. This means that the lower perceived ease of use did not influence the positive perception of user engagement and interactivity. When looking at the interview results, one of the main issues of crowd-feedback tasks was that crowdworkers were insecure about the focus of the study, the right specificity of their feedback, and had problems with keeping an overview of their feedback. As the category feature addressed all of these problems, crowdworkers felt more secure and used the categories as guidance for the task, which might have covered up the usability issues and in turn led to the high value for perceived ease of use. Regarding the importance of features for feedback, crowdworkers still ranked star ratings and markers higher. While ratings and markers enable feedback providers to enrich their textual feedback with additional feedback, the categories only offer a better structure. For markers, the crowdworkers' experience with the feature matches the quantitative outcomes as the feedback got more specific in this treatment. Comparing the star rating treatment that contained also categories with the treatment with only categories, the user engagement was lower while the feedback quality was higher. The user engagement was even the lowest for the star rating treatment. The reason for this might be the increased complexity of two features that lead to a higher mental workload for feedback providers. Feedback providers ranked the speech-to-text feature as the least important which is consistent with them not using it at all. The interviews and the feature ranking confirmed that they perceive the feature as a nice add-on, but not essential for providing good feedback. Still, the pure presence of the feature had a positive impact on crowdworkers' perceptions and the resulting feedback compared to not having

any feature included. In the interviews, workers were not completely averse to using the feature. We assume after workers get familiar with feedback-providing tasks, some will start using the feature. Still, the value and effect of the speech-to-text feature should be analyzed in future studies.

3.5.3. Design Implications

Based on our results we here provide a summary of design implications for crowd-feedback systems.

Focus on Crowdworkers' Perceptions

Crowdworkers perceived the scenario feature as the most important. Therefore, we recommend providing a scenario in feedback tasks to guide feedback providers. Although the category treatment performed the best in terms of user engagement, markers, and star ratings were perceived as more important by crowdworkers. This is consistent with the qualitative results as many workers who did not have a marker feature in their version of *Feeasy*, asked for a feature to annotate the user interface of the design prototype. For the star rating, this does not apply. Consequently, markers seem to have a bigger positive impact on crowdworkers perceptions of the crowd-feedback system and should therefore be applied additionally to the scenario when aiming to positively impact crowd perceptions.

Focus on Feedback Quality

The feedback quality was the lowest for the full and the basic treatment. Consequently, we recommend applying selected features when designing crowd-feedback systems and paying attention to balancing the advantages of multiple features and the increased complexity. For the single-feature treatments, there is no feature that clearly performed better than the others. Each feature has individual advantages and feedback requesters must understand their feedback requirements to select the appropriate features.

Focus on Feedback Quantity

When aiming for many feedback comments, feedback requesters should apply categories or markers. Adding star ratings to the categories has only a minor negative impact on the feedback quantity and could therefore also be an option. Regarding the length of feedback comments, scenarios are the best choice, followed by categories (with and without star

ratings), and markers. Applying all five design features has a negative impact on the length of feedback. As each feature takes some time to get familiar with it, generally fewer features are beneficial when aiming for many long feedback comments. Regarding the number of comments, categories are the favorable design features.

3.6. Limitations and Future Work

While we followed a rigorous evaluation approach several limitations apply to our study. In the following, we provide an overview of limitations and present future research directions.

Relevance of Design Prototypes

First, our crowd-feedback artifact *Feeasy* was designed for the collection of feedback for all sorts of design prototypes over all phases of the design process. However, in our evaluation studies, we used always the same design prototype to guarantee for comparability of the results. This was necessary since we focused on the evaluation of the design features. Future work should expand the design feature evaluation with additional design prototypes from different design phases. As we learned that workers use their personal expectations and experiences with similar websites to come up with valuable feedback, the workers' requirements for a crowd-feedback system could be much different when the feedback is collected on a less common type of software.

Investigation of the Speech-to-Text Feature

In all of our three studies, no worker has used the speech-to-text feature. Consequently, the reported perceptions and effects on the feedback are only based on workers' assumptions about their interaction with the feature. Additionally, the changes in the feedback quality and quantity are only caused by the presence of the feature. On the one hand, this shows that the sole presence of features has an effect on crowdworkers. On the other hand, we are not able to make statements about how the usage of a speech-to-text feature affects feedback quality and quantity. In our studies, workers reported multiple advantages and disadvantages of a speech-to-text feature. Especially a recording feature would enable designers to consider more factors than just the pure content of the feedback (e.g., tone). Therefore, we suggest future research to study voice input features for feedback individually. The results might also be relevant for other domains like app store reviews.

Interdependencies between Design Features

In this paper, we presented two studies focusing on a specific set of individual design features relevant to crowd-feedback systems. We assume that there exist interdependencies between the individual design features. Especially in the first design study, the perceptions of the participants might be influenced by interdependencies of the individual features. We attempted to counteract this by asking participants specifically about their perceptions of each individual feature.

Still, the analysis of potential interdependencies is beyond the scope of our paper. However, understanding how the combination of design features affects crowdworkers' perceptions and the resulting feedback, might be very valuable for the design of crowd-feedback systems in research and practice. Therefore, future research should expand on an analysis of the interaction effects of design features. This knowledge can be used by feedback requesters (i.e., designers) to design crowd-feedback systems according to the requirements of their feedback studies. To enable feedback requesters to instantiate these individualized crowd-feedback systems without large effort, a crowd-feedback system configurator would be beneficial. This configurator could guide feedback requesters in creating dedicated crowd-feedback systems that are adapted to their needs. This would enable designers and developers to easily integrate crowd-feedback systems in all phases of their software lifecycle.

3.7. Conclusion

Design features of crowd-feedback systems have an impact on the resulting feedback. While most existing studies in this context focused on analyzing the feedback outcomes for requesters, we aimed to align crowdworkers' perceptions on a spectrum of different design features with quantifiable effects on feedback quality and quantity. We conducted two studies, in which we first developed the crowd-feedback system *Feeasy* and, subsequently, used it to analyze distinct five design features for crowd-feedback systems. Our results provide evidence that more design features are not beneficial in all use cases, but applying any design features is better than none. Furthermore, we learned that scenarios and markers are favorable design features when considering the crowdworker perspective, while for the feedback quality and quantity, it is primarily important to not overwhelm crowdworkers with too many complex features. Still, the application of any feature improves feedback

quality and quantity. We enrich these findings with profound details on the advantages and disadvantages of each design feature as perceived by crowdworkers. Our findings motivate further investigations for the future design and configuration of design features which are combined to achieve specific effects and serve as a basis for the development of a crowd-feedback system configurator. Overall, we contribute with our work to make the software development process not only more scalable but also more human-centered.

4. Study III: Designing Configuration Systems for Crowd-Feedback Request Generation

4.1. Introduction

Continuous user involvement in the software evaluation processes is crucial for the success of software development and one key activity in the user-centered design process to ensure the fulfillment of functional and non-functional requirements. The drawbacks of the user-centered evaluation of software are high costs and scalability issues (Scholtz, 2001). A solution for these challenges is crowd-feedback systems. These systems leverage crowdsourcing to collect large amounts of structured design feedback (Wauck et al., 2017). The focus of crowd-feedback systems is to collect explicit feedback on the perception of graphic designs (e.g., posters) (Xu & Bailey, 2014), interactive prototypes (Oppenlaender, Kuosmanen, et al., 2021), websites (Oppenlaender, Tiropanis, & Hosio, 2020) or other types of software (e.g., chatbots) (Choi et al., 2021). Research has shown that crowd-feedback systems can not only enable scalability and reduce costs as well as effort for feedback requesters (Dow et al., 2013; Oppenlaender, Kuosmanen, et al., 2021; Yen, Dow, et al., 2016) but can also produce reliable feedback that helps to improve the resulting designs (Yuan et al., 2016). However, crowd-feedback systems do not receive much adoption in practice. There are several potential reasons for this. Crowd-feedback systems are usually fixed to specific use cases and provide limited flexibility for feedback requesters to adapt the system to their needs (Haug & Maedche, 2021a). Especially designers with no development skills or limited methodological knowledge might have challenges applying crowd-feedback systems in practice. When creating crowd-feedback requests, many decisions need to be taken. Feedback requesters must not only decide which crowd to ask for feedback and how to incentivize it, but also what type of feedback is required (qualitative vs. quantitative), on which aspects the feedback is required (functional attributes vs. non-functional attributes vs. content), and how feedback providers shall be able to share their feedback (e.g., voice recording, collaboration, or markers) (Haug & Maedche, 2021a). This determines what features of the crowd-feedback system shall be leveraged and in which way they shall be combined. The mentioned challenges represent a large obstacle to the adoption of crowd-feedback systems and hinder the sufficient use of such systems. The

research field of end-user development extensively explored the need for tailoring software systems to individual requirements by domain experts without programming skills. The goal is to make systems not only easy to use but also easy to develop (Lieberman et al., 2006). Existing research proposed design knowledge for configuration systems, however, mainly with a focus on the manufacturing or production domains (Anish & Ghaisas, 2014). Research on configuration systems to create individualized software is rather scarce. Thus, there is a need to investigate the specific context of designing configuration systems to enable feedback requesters to derive successful software evaluation strategies. Consequently, we articulate the following research question (RQ): *How to design a configuration system to support designers in creating effective customized crowd-feedback requests?*

To answer this RQ, we, first, conducted 14 qualitative interviews with design experts to understand current issues regarding software evaluation. These interviews provided us with initial insights that we confirmed and extended with a literature review. Based on these insights, we developed four design rationales for a crowd-feedback request configuration system. We then instantiated them in a configuration system for crowd-feedback requests for customized software evaluation. Finally, we evaluated this artifact in an exploratory focus group workshop. With our work, we contribute to research by

- First, providing an understanding of designers' challenges of software evaluation.
- Second, proposing and evaluating four design rationales for the design of a configuration system for crowd-feedback requests.
- Third, developing a configuration system for an existing crowd-feedback system based on the proposed design rationales.

Individual design feedback requests will make crowd feedback applicable to a more diverse set of use cases and increase the integration of software evaluation in software development processes.

4.2. Conceptual Foundations and Related Work

In this section, we provide conceptual foundations on design feedback in general and crowd-feedback systems in particular. Further, we provide an overview of related work on configuration systems.

4.2.1. Design Evaluation & Feedback

The evaluation of designs is the fourth step in the user-centered design (UCD) process (ISO 9241-210, 2019). The goal of the evaluation phase is to iteratively refine the design until the users' needs are met (Brhel et al., 2015). Traditional evaluation methods that build upon explicit user involvement include but are not limited to usability tests, interviews, and focus groups (Brhel et al., 2015). These methods are often fraught with scalability issues (Scholtz, 2001). This general problem illustrates why continuous evaluation of designs with user involvement is often a major challenge in UCD (Brhel et al., 2015). In addition, designers often lack the required methodological knowledge to properly involve users in the evaluation process (J. Y. Mao et al., 2005) and do not have access to a diverse set of (potential) users (Ma et al., 2015). Running software is usually evaluated via feedback pop-ups or in dedicated feedback forums (Almaliki et al., 2014). However, these approaches usually generate little feedback with low quality due to a lack of focus and structure (Almaliki et al., 2014). Consequently, in practice, only a small percentage of development projects engage with users in every stage of the UCD approach (J. Y. Mao et al., 2005). To overcome some of these issues, crowd feedback has emerged as a new approach to collecting design feedback in a more scalable way. Crowd feedback originates in the graphic design domain as an alternative to peer feedback (Yuan et al., 2016). Dedicated crowd-feedback systems offer feedback providers structure that increases the feedback quality (Luther, Tolentino, et al., 2015). Another benefit of crowd feedback is that designers do not necessarily need access to real users anymore but can also use an anonymous crowd as it can be found on platforms like Amazon Mechanical Turk (MTurk). Still, crowd feedback can also be collected from stakeholders or potential users (Easterday et al., 2017; Oppenlaender, Tiropanis, & Hosio, 2020). In the next section, we present what dedicated crowd-feedback systems look like and what new challenges these systems bear.

4.2.2. Crowd-Feedback Systems

Crowd-feedback systems present an approach to solving the scalability issues related to user involvement in software development. These systems collect large amounts of structured feedback by engaging a group of humans, which can be but not must be real or potential users (Haug & Maedche, 2021a). In research, various studies on crowd-feedback systems exist. A big advantage compared to simple surveys as they can be created with tools like

LimeSurvey or Google Forms is that crowd-feedback systems usually enable the feedback collection while the feedback provider is actually using the system (e.g., Xu, Huang, et al., 2014; Haug, Benke, Fischer, and Maedche, 2023). Crowd-feedback systems can also apply dedicated features that are not available in online survey tools like markers to pin comments to a specific element (Haug, Benke, & Maedche, 2023). According to Haug and Maedche (2021a), the core element of crowd-feedback systems is their design characteristics and the selected crowd configuration. The crowd configuration describes what type of crowd is asked for feedback (anonymous, users, students, and convenience) and how this crowd is incentivized (money, involvement and improvement, credits, social compensation, and gamification). The design characteristics are split into feedback collection mechanisms and interactivity cues. While feedback collection mechanisms (questionnaire, free text field, categories, selection, and direct manipulation) conceptualize all features to collect feedback, the interactivity cues (collaboration, markers, context, and recording) describe additional features that help feedback providers to improve their feedback quality or enrich their feedback (Haug & Maedche, 2021a). While researchers have demonstrated that crowdsourcing high-quality feedback is feasible with dedicated crowd-feedback systems, there are still discussions about definitions and measures for design feedback quality (Haug, Benke, Fischer, & Maedche, 2023). Also, feedback might be less honest when people are paid for providing feedback (Haug, Benke, Fischer, & Maedche, 2023).

Most of the existing crowd-feedback systems have focused on evaluating static designs like posters. For example, *Voyant* is a popular example of a crowd-feedback system with the goal of collecting feedback on graphic designs (Xu & Bailey, 2014; Xu, Huang, et al., 2014). It captures the crowd's first impressions and how well specific goals and design guidelines are met. There also exist a few systems that evaluate interactive designs like chatbots, mock-ups, or running websites. Many of the studies focused on achieving a high feedback quality or optimizing the resulting designs. Thereby, the applicability of these systems in practice was often neglected (Haug, Benke, & Maedche, 2023). Consequently, all existing crowd-feedback systems in research are fixed to a specific use case or to the evaluation of a specific software system (Haug & Maedche, 2021a). One system that allows the evaluation of interactive designs and combines multiple design characteristics of crowd-feedback systems to understand how they are perceived by users is *Feeasy* (Haug, Benke, & Maedche, 2023). *Feeasy* collects feedback via a free text field, categories, and

star ratings that are attached to the categories. Additionally, it contains markers, context, and recording, in the form of a speech-to-text feature as interactivity cues. Haug, Benke, and Maedche (2023) used *Feeasy* to understand the effect of the design characteristics of crowd-feedback systems on the resulting feedback quality and quantity. Their results can serve as a basis to better support designers in creating individual crowd-feedback requests and tailoring the design to the respective requirements and context. For an extensive overview of existing crowd-feedback systems, we recommend the literature review of Haug and Maedche (2021a).

4.2.3. End-User Development and Configuration Systems

End-user development (EUD) shall empower software users without a background in programming to develop or modify their own software systems (Lieberman et al., 2006). This allows for more flexible and tailored use of these applications. Research in the field of EUD proposes methods, techniques, and tools for creating, modifying or extending software artifacts (Lieberman et al., 2006). As in our case, the designers who shall configure the crowd-feedback systems are not the actual end-users of the crowd-feedback system, we will consider EUD only as a side topic in the design of our configuration system.

Regarding configuration systems, there mainly exist two types: needs-based and parameter-based (Randall et al., 2007). Needs-based configuration systems ask users to specify the relative importance of their needs regarding the resulting product. An algorithm then combines the design parameters so that the user's needs are matched as closely as possible. Parameter-based configuration systems on the other hand allow users to directly specify the design parameters for the resulting product. Therefore, these systems are usually more flexible but also require more expertise from users (Randall et al., 2007). Research on configuration systems for adapting software to users' needs is rather scarce. Feine et al. (2019) designed a chatbot social cue configuration system. This system supports chatbot engineers in accessing descriptive knowledge to make more justified social cue design decisions by transforming the descriptive knowledge into prescriptive knowledge. While this configuration system is applying a needs-based approach by providing recommendations based on the target user, task, and context, we will aim for a parameter-based approach. This is because the descriptive knowledge of design features of crowd-feedback systems and their effects is rather limited and therefore no justified design decision can be auto-

matically derived purely based on users' needs. Parameter-based configuration systems are characterized by many decisions that users must make. So, on a more abstract level, the design of crowd-feedback requests can be seen as a series of consecutive decision tasks in which the feedback requester is the advice taker, and the configuration system serves as the advice giver. In this analogy, selecting a feature or a type of crowd can be seen as a single decision task.

4.3. Designing a Configuration System for Feedback Request Creation

To design a configuration system that can effectively support designers in creating individual crowd-feedback requests, we combined insights from expert interviews with an exploratory literature review. In the first step, we focused on collecting the fundamental requirements of designers and product owners and understanding how current solutions need to be improved to meet these requirements to develop design rationales for our system design. In the next step, we developed a software prototype based on these design rationales.

4.3.1. Interview Study and Literature Review

We interviewed fourteen design experts to understand what issues they experience when evaluating prototypes or software and what the related processes look like. The design experts (64.29% female) included UX designers, UX managers, UX researchers, and product owners. The interviewees were on average 36.29 years old (SD = 9.96 years), mainly worked in large companies, and had on average 10.71 years (SD = 8.03 years) of work experience. The interviews took on average 24.51 min (SD = 4.92 min). The interviews were conducted in German, then transcribed, translated, and coded following an empirical-to-conceptual approach, mainly focusing on designers' issues related to design evaluation. To verify our results and to get a broader picture of designers' challenges when evaluating designs and potential solutions, we conducted an exploratory literature review on existing crowd-feedback systems and the related feedback processes. We used publications from a literature review on the state-of-the-art of crowd-feedback systems by Haug and Maedche (2021a) as a starting point and extended our search based on their papers. To identify

and structure the issues, we coded all 21 papers that we found relevant following an iterative empirical-to-conceptual approach (Nickerson et al., 2013). This also helped us to assess the importance of identified issues based on their number of occurrences in different papers. In the following, we bring together the insights from the interviews and learnings from the literature.

Regarding the implementation of software evaluation practices, we identified four core problem areas based on our interviews: budget and time, process and methodology, internal collaboration, and diversity of participants. Exemplary quotes for these problems are summarized in Table 4.1.

Table 4.1.: Results of expert interviews

Problems with design evaluation in practice	**Quotes**
Budget and time	*"Very often, strict evaluation is simply dispensed with because of the costs involved, and people say they'll go live and try to fix things with the live feedback." (E.1)*
	"That is also the biggest pain point of my colleague or of the UX people [...], that in many projects this phase comes too short. Be it budget or be it time." (E.3)
	"Of course, that would be very nice if you still had the time and budget to really involve the user in the entire process. Well, that's often simply not possible, also from the client's point of view." (E.1)
Process and methodology	*"That's actually always the challenge in the enterprise setting, or in the business UX field, that you have the right people in place at the right time." (E.6)*
	"This [the evaluation process] is relatively unstructured, it is not like we have a questionnaire or somehow collect structural user feedback." (E.9)
Internal collaboration	*"Because you are somewhere dependent on the product owner, you can't go out on the street yourself, you have to coordinate with the product owner somehow and often I have the feeling that this is not so important to them because they always say that they know what the end user needs." (E.3)*
Diversity of participants	*"If I do then I might ask 3-4 people and the feedback on the prototypes is then based on the feedback from those three to four people." (E.11)*
	"And that's where I definitely still see a big gap [...] that we still have a lot of only internal feedback and, above all, only feedback from people who are involved there anyway." (E.9)

When looking at the literature, we learned that many of these challenges could be solved by applying crowd-feedback systems. Crowd feedback offers a scalable approach to collecting feedback from a diverse group of people. Accordingly, we wondered, why crowd feedback is not applied in practice and took a closer look at the challenges that potential feedback requesters might face. From the analysis of existing literature, we learned that the preparation of crowd-feedback requests is experienced as time-consuming and effort-intensive (Hui et al., 2014; Snijders et al., 2015). Especially for requesters with little technical experience

and skills, the creation of a crowd-feedback request is a complex challenge (Oppenlaender, Kuosmanen, et al., 2021). Requesters with little methodological knowledge also find it difficult to learn the necessary techniques and skills to create a feedback request (Hui et al., 2014; J. Y. Mao et al., 2005).

Decisions on crowd-feedback system features and settings are complex and hard to make without specific knowledge. For example, research shows that the choice of the crowdsourcing platform that is used for the crowd-feedback request affects the received feedback (Yen, Dow, et al., 2016). Paid task markets are found to provide feedback with more design suggestions while responses from web forums lead to more process-oriented feedback (Yen, Dow, et al., 2016). Also, most incentives are linked to the type of crowd used. For example, the anonymous crowd is found to be mostly financially incentivized (Haug & Maedche, 2021a). Also, the features of crowd-feedback systems need to be adapted to the goal of the feedback study and to the perceptions of feedback providers (Haug, Benke, & Maedche, 2023). Supplementary, one must consider that different end-users have different design needs (Luther, Tolentino, et al., 2015). For instance, the support for structured feedback can be reduced as soon as the crowd users gain some experience (Oppenlaender, Tiropanis, & Hosio, 2020). Existing crowd-feedback systems are found to have a fixed set of design characteristics (Haug & Maedche, 2021a). Consequently, they are fixed to one specific evaluation and are not adaptable to other use cases (Luther, Tolentino, et al., 2015). The inflexibility of crowd-feedback systems is further illustrated by the findings that they are mostly fixed to the evaluation either during the development or operation phases (Haug & Maedche, 2021a). Only a few systems focus on the evaluation of interactive systems (Haug & Maedche, 2021a). This means that existing crowd-feedback systems have usually a fixed set of design characteristics and can hardly be reused for other use cases.

Based on these insights, our goal was to develop a configuration system that shall help design experts without knowledge of crowd feedback to create individual crowd-feedback requests for their design projects.

4.3.2. Design Rationales

With respect to the goal of our study, we synthesized the findings of our initial interviews and literature review to develop four design rationales (DRs) for our system design:

1. *User Guidance.* Users of our system might have varying methodological knowledge and technical skills. Therefore, our system needs appropriate functions to guide and support feedback requesters during the feedback request creation process so that they can quickly and easily create a crowd-feedback request. This includes the selection of crowd-feedback system features, but also the distribution of feedback requests.

2. *Effect of Design Characteristics.* For designers without any experience in crowd feedback, it is not possible to comprehend the effects of individual design decisions regarding the feedback request. Our system shall display the potential effects of feature selections on the resulting feedback so that feedback requesters can select and combine appropriate features according to their current requirements.

3. *System Customization.* Every design project and therefore every feedback request is individual. Our system needs functions to customize feedback requests so that feedback requesters can collect feedback for different tasks and use cases.

4. *Crowdworker Perspective.* We want crowdworkers to be able to submit high-quality feedback when they are using the configured crowd-feedback system. Therefore, our configuration system needs to have functionalities that allow the feedback requester to consider the crowd's needs and requirements during the configuration process.

4.3.3. System Design

We developed a prototype based on our four design rationales. For this, we combined a real software artifact with a Figma design mock-up. This allows us to reduce the development effort before verifying the design rationales are valid and to quickly iterate the design. The selection process of design features of the configuration is thereby implemented in the software prototype, while the platform which feedback requesters can use to manage their requests is implemented as a design mock-up in Figma. We aimed to keep our design of the configuration system very flexible, so that it can be used for a diverse set of designs including interactive prototypes or mock-ups, but also static designs like posters. Figure 4.1 shows our software artifact and a final feedback request. The colored boxes indicate how our four design rationales were transferred to design features. Feedback requesters start in the *Projects* tab in the configuration system. There, they can manage their feedback projects and related feedback requests. Additionally, the platform contains an *Apps* tab that presents different crowdsourcing platforms and their characteristics, including the

advantages and disadvantages of each platform, and supports designers in deciding how to distribute their feedback requests. The third tab in the menu is the *Help Center*. There, users can get support for interacting with the configuration system, as well as learn about descriptive knowledge for feedback requests in the form of guidelines. For each design project, users must upload their design in the form of one or multiple files. These files can be images for static designs or multiple HTML files for interactive designs. Then, they can create a new feedback request. The step-by-step configuration of the feedback request is implemented as a software artifact.

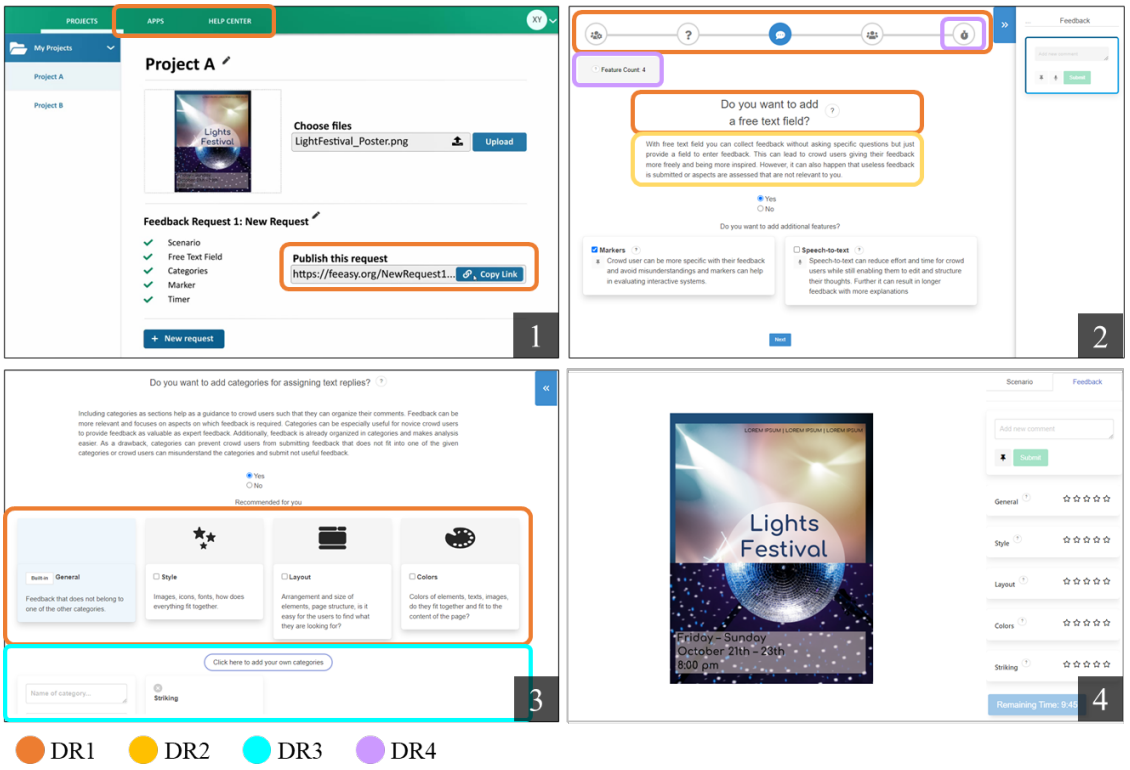


Figure 4.1.: Configuration system artifact as an instantiation of our four design rationales. (1) For each project, users must add their design via a file upload. (2) When creating a new feedback request users are guided through each step. On top they have a progress bar with five steps, on the right side, they see how each feature would be implemented in their feedback request. (3) After deciding on a feature, users can configure the feature. (4) The final feedback request shows the uploaded design and the feedback panel.

Based on the core features of *Feeasy* (Haug & Maedche, 2021b) and additional feedback features according to Haug and Maedche (2021a), users are guided through five configuration steps as displayed in Figure 4.2. First, feedback requesters can choose if they want to offer feedback providers a context in the form of a scenario and add text to each step in the scenario. This can help feedback providers especially, when feedback is collected on an

interactive prototype to navigate through it. Then, they can choose if they want to include a questionnaire either with answers in the form of scales or with text entries. The third step allows them to add a free text field and related features, such as categories with a star rating, markers, and a recording feature. In the fourth step, users can decide if they want to include a collaboration feature that enables feedback providers to see the comments of others and react to them. Finally, feedback requesters can decide if they want to give users a predefined time after which they are allowed to submit their feedback. During the whole configuration process, a panel on the right side of the screen shows users how implementing the respective feature would look. For each design option, a tooltip with a definition of the feature is next to the question and a box with descriptive knowledge of the design option is below the headline. For some features, the configuration system offers recommendations, e.g., categories that are frequently used in research. Feedback requesters are also able to customize the feedback features, for example, add individual categories. The user interface for the configuration process also contains a feature count that shows users how many features they have already combined. This shall help them to keep the perspective of the feedback providers in mind and not overload the feedback request with unnecessary features. When feedback requesters are done with their configuration process, they get redirected to the projects tab on the configuration platform. There, a link is created that leads feedback requesters to the configured feedback request. The idea is that feedback requesters can then share this link with their desired crowd or integrate it into a survey.

4.4. Evaluation Study

To confirm our design rationales and understand how potential feedback requesters perceive our configuration system, we conducted an evaluation in the form of two exploratory focus group workshops. We decided on an in-person workshop instead of following a crowdsourcing approach as our goal was to directly interact with potential users and be able to react to their feedback, questions, and ideas immediately. In the following, we present the evaluation procedure and describe our participants.

4.4.1. Procedure

Our configuration system shall work for all types of graphic and interactive designs. Therefore, we conducted two workshops. While the first workshop was focusing more on evalu-

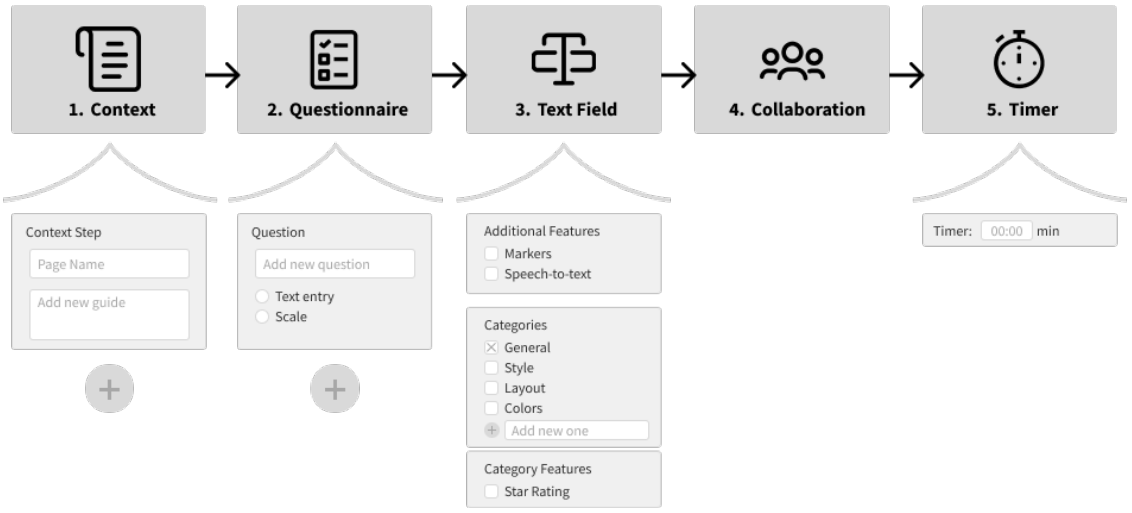


Figure 4.2.: Configuration process as it is implemented in our instantiation of the configuration system. First, designers can add context to their design. In multiple steps, they can explain the design and guide feedback providers through the interaction with it. In the second step, designers can add a questionnaire and enter questions that can be answered via text entry or a 5-point Likert scale. If designers decide to add a free text field in the third step, they can also choose additional features, such as markers, speech-to-text functionality, categories, and star ratings. The fourth step allows designers to add a collaboration feature. And finally, in the fifth step, designers can add a timer to limit the time users have to provide feedback.

ating interactive designs, such as website prototypes, the second one addressed the evaluation of static designs like posters or simple mock-ups. After a short introduction to crowd-feedback systems and crowd-feedback requests in general, we provided the participants with an overview of the steps in our configuration system. We then showed them an exemplary configuration for creating a feedback request for the evaluation of a simple website prototype (workshop 1) or a poster design (workshop 2). The demonstrating researcher presented the use case and showed which decisions a feedback request creator must make. After the demonstration, the open questions of the participants were clarified. Then, each participant had ten minutes to try out the configuration system him/herself. In the next step, we explained the Strengths-Weaknesses-Opportunities-Threats (SWOT) analysis method to the participants that we used to structure the exploratory focus group. The participants then had time to write down their perceived strengths, weaknesses, opportunities, and threats of the configuration system. Finally, the results were read out loud and explained by the participants. Other participants and the researchers could discuss ideas or ask follow-up questions to clarify each point. In the second workshop, the moderating researcher presented the results of the first workshop, initiating a discussion

about the similarities and differences between the results. With this input, participants had the opportunity to expand their SWOT analysis.

Both workshops were recorded and the final results in the SWOT matrix were captured. As the workshops were conducted in German, the recordings were first transcribed and then translated into English. Then, one of the authors analyzed and coded the SWOT results and transcriptions regarding the four design rationales.

4.4.2. Participants

Exploratory focus groups propose improvements to refine the design (Tremblay et al., 2010). As our configuration system shall be designed for different user types and the evaluation of different types of designs, we aimed for a diverse set of participants. This is also consistent with the recommendation to mix different skill sets for the evaluation of decision-aid tools (Tremblay et al., 2010). Focus groups should involve between four and twelve participants (Morgan, 1997). Therefore, we invited ten participants with various backgrounds and levels of experience to participate in our evaluation (nine male, MAge = 31.1 years, MDesignExp = 8.4 years, MDesignEvalExp = 4.8 years) and split them according to their design background in the two workshops: P.1 to P.6 participated in Workshop 1 and P.7 to P.10 in Workshop 2. Details on our participants are shown in Table 4.2.

Table 4.2.: Demographic information on participants of the focus group workshops

ID	Profession	Gender	Age	Experience in years		Level of Design
				Design	Design Evaluation	Knowledge
P.1	Experience Manager	M	39	20	20	Very high
P.2	UX Designer	F	24	2	3	Medium
P.3	IT Specialist	M	44	12	8	Medium
P.4	IS Student	M	26	0	0	Very low
P.5	Web Designer	M	28	8	7	High
P.6	UX Developer	M	29	4	1	Medium
P.7	Graphic Designer	M	42	23	0	Very high
P.8	Graphic Designer	M	27	7	3	High
P.9	PhD Student in IS/HCI	M	28	4	2	High
P.10	Media Communications Designer	M	24	4	4	High

4.5. Results

We analyzed the results of the SWOT analysis according to our four design rationales. The most important results for each design rationale are summarized in Table 4.3 and explained in more detail below.

Table 4.3.: Summary of the results of the SWOT analysis according to design rationales (DRs)

DR	Strengths	DR	Weaknesses
S1.	Feedback requesters are guided through the configuration process simply and efficiently.	W1.	Some parts of the configuration process lack clarity and are very complex.
S2.	The effects of different design options are sufficiently explained, especially when users are inexperienced or do not know their goal.	W2.	The learning effects of users are not considered in the design.
S3.	The feedback request is very flexible and can be adapted to different use cases.	W3.	The configuration system and the options are at some points too open.
S4.	The perspective of the crowd is always in view via the feature count.	W4.	The connection between decisions and the meaning regarding the crowd's perception was sometimes not obvious.
DR	**Opportunities**	**DR**	**Threats**
O1.	Offering templates could simplify and speed up the configuration process.	T1.	The incentivization for feedback providers should not be neglected.
O2.	Including experts' recommendations would help in making the right configuration decisions.	T2.	Users could get lost in the flexibility of the system.
O3.	The configuration system could offer even more design features.		

In general, our participants appreciated the user guidance provided by the configuration system. This was explained by P.6 as follows: *"I think the user guidance is very well done. I had the feeling that even though I'm seeing it for the first time, I always know where I am, what I'm doing, how to continue, how to get back"*. In addition, participants expressed that the system is *"fast and simple"* (P.8) and allows them to quickly create a feedback request (*"I could quickly upload every little piece and get rapid feedback to evaluate it the next day"*, P.8). However, at the same time, the configuration system was still perceived as very complex and *"... the look and clarity of what exactly is being applied could be slightly [...] enhanced from the user guidance perspective"* (P.5). One reason for the high perceived complexity is that *"... you are flooded with a lot of text"* (P.10). P.9 explicates: *"Quite a lot of text at once and the text itself not that appealing to read"*. Therefore, participants recommended *"... hide out the points that you might understand, so you can sort of pop*

in what you are interested in to get more information, in a popup window for example" (P.10). Further, participants would have liked to receive more guidance in selecting the right crowd and appropriate incentive as they perceive this as a major problem in feedback requests. The integration of research insights into the configurator was perceived as a major strength of our configuration system: *"Particularly for inexperienced users, I found it good that the positive and negative effects of features were described in the description"* (P.4). Concerning the explanations in the information text on the design features, it was noted that a learning effect of the users is not taken into account. Some information is only relevant for the first usage, *"... which you would then probably not require the second time anyway"* (P.8). Participants proposed to include even more recommendations, also based on design experts' experiences. Regarding our third design rationale, the configurability of the feedback requests, participants were indecisive if they liked the flexibility and the openness or if it was too much for them: *"It [the configuration system] seems to be flexible. I had the feeling that I could go into a lot of detail and do a lot of things. On the one hand, I think that's good, but on the other hand, it's not good"* (P.1). Some participants requested having even more design features included that they could configure in their feedback requests. Lastly, we integrated the perceptions of the crowd in the form of the feature count, this led to participants always having the perception of the crowd in their view. One participant reflected on this as follows: *"I think the count of features at the end is pretty cool. Because you can see if it becomes too crowded when you give feedback. I think that makes sense"* (P.9). However, it was not clear to all participants how to consider the crowd's perception in the form of the feature count in their feedback request. One participant stated that s/he wasn't sure *"...what that meant for my specific request. So, I wasn't clear on whether I did it well or not"* (P.2).

4.6. Discussion

While crowd-feedback systems offer feedback requesters a scalable and effective way to collect feedback on graphic designs, prototypes, and software designs, existing crowd-feedback systems do not support feedback requesters in creating and configuring individual feedback requests. To address this problem, we designed a configuration system that supports feedback requesters in individual crowd-feedback request creation. Drawing on existing research on challenges in design evaluation, the effects of feedback features, and

decisional guidance, we examined how the combination of descriptive knowledge and a step-by-step configuration process can support feedback requesters in creating individual feedback requests. Subsequently, we conducted an exploratory focus group evaluation with a diverse group of designers. The results of our evaluation show that our approach has the potential to solve existing problems of designers, while there are still more ideas that could be explored in future research. We learned that offering too many design options makes designers feel lost in the process. Also, they emphasized the idea of offering templates based on experts' recommendations. Therefore, our study provides valuable theoretical contributions and practical implications that we discuss in the following.

4.6.1. The crowds' perspective still needs more attention

We did not receive as much feedback on DR4 as on the other three DRs. This might be caused by feedback requesters still not having the crowd's perceptions in mind when creating feedback requests. We instantiated this rationale mainly via the feature count. One main issue with the feature count was that participants felt that it was not integrated sufficiently into the prior decisions. It was not clear enough how users shall consider the feature count when making their design decisions. Further, we interpret the lack of feedback on DR4 so that feedback requesters do not think about the crowd's perspective much as they assume that the configuration system automatically respects their needs. Consequently, in further iterations, we need to come up with alternative instantiations of DR4 to ensure the crowd's perception is always considered during the configuration process.

4.6.2. There is a trade-off between flexibility and complexity of the configuration system

Two core topics during the focus group evaluation were the flexibility of the configuration system due to the combination of many feedback features and the high complexity of the configuration process due to much information and many interdependencies that must be considered. This issue demonstrates tensions between our design rationales, especially DR1 and DR3. On the one hand, feedback requesters need guidance, on the other hand, they want a flexible configuration system so that they can completely customize each feedback request. We believe that every configuration system needs to find a balance between flexibility and complexity. Although both aspects are not two extremes on a

shared scale, they need dedicated features to be compatible as often one comes at the price of the other. While needs-based configuration systems are less complex for users, they provide less flexibility and transparency (Randall et al., 2007). The lack of transparency was also a major concern in the evaluation of the configuration system of Feine et al. (2019). We wanted to counteract this problem by offering a parameter-based configuration system. This was also necessary because there is still a lack of knowledge of the effects of feedback features. Crowd-feedback systems are complex systems with multiple features to combine. To reduce the complexity of our configuration system, many participants suggested offering templates for specific use cases that can then be adapted by feedback requesters. This would be a hybrid approach of the needs-based and parameter-based systems. These templates could be based on the three research streams identified by Haug and Maedche (2021a): anonymous crowd-feedback, real user crowd-feedback, and hybrid crowd-feedback. By offering a template, users would need to make one decision when choosing a template and can then make additional fine-tuning decisions but they are not forced anymore to decide for or against every single feedback feature. The templates could be developed by experts considering theoretical knowledge and practical implications of feedback features. The templating approach is consistent with the recommendation of Weinmann et al. (2011) who suggested offering a hybrid approach for configuration systems, also for users with different levels of expertise. An alternative to offering templates by the system could be allowing users to share their requests and reuse feedback requests of other users. By rating or commenting on feedback requests of others, templates could be created organically by the users of the configuration system. This would also bring the system closer to the idea of end-user development that we mentioned earlier.

When offering templates, another approach could be to implement them in established survey tools (e.g., LimeSurvey) similarily to *QButterfly* (Ebert et al., 2023), a toolkit for conducting usability studies in LimeSurvey or Qualtrics. Thereby, we can achieve the same benefits: reduce authoring time and complexity, empower users without programming skills to conduct design studies facilitate the re-use of the existing functionality of these tools, and facilitate the replication of ideas.

4.6.3. Experienced users of the configuration system might need advanced functionalities

We explicitly decided to design our configuration system for both, novice and expert feedback requesters. Therefore, some features of the configuration systems might be more useful for novices, while others are specifically designed for experts. Parameter-based configuration systems need users usually to be experienced in the specific domain to make the right decisions (Randall et al., 2007). As we did not assume that all users already know which are the right decisions for their use case, we focused much on user guidance by explaining the advantages and disadvantages of feedback features. However, one thing we did not consider sufficiently here was the learning effects of users. This point was raised during the evaluation of our configuration system. When users have understood the effects of the design features, they do not need to read the information on advantages and disadvantages again every time they want to create a new feedback request. Therefore, participants requested to adapt the UI more to experienced users of the configuration system by allowing them to hide information texts or store user inputs for the following configuration processes.

4.6.4. Limitations and Future Research

Of course, there are also limitations in our work that need to be considered. First, participants in our focus group followed an artificial use case when they tested the prototype. These limitations might have biased participants' perceptions of our system. In future work, we want to develop a completely functional artifact. This will make the evaluation results even more insightful and reliable. Second, we used an exploratory focus group to perform a qualitative evaluation of the configuration system artifact. The goal of the evaluation was to understand if the configuration system actually supports feedback requesters in creating design feedback requests and how we can further improve the system. While this approach enabled us to collect valuable insights into users' interaction with the configuration system and innovative ideas, future research should conduct a quantitative evaluation to understand to what extent the configuration system is usable and helpful for feedback requesters.

Overall, our design rationales showed to be key for providing feedback requesters the possibility to create individual crowd-feedback requests. They are partly contradictory,

which means that a good balance between them, especially the flexibility and complexity of the configuration system, needs to be found. Our evaluation also sheds light on additional design issues, which offer valuable starting points for further improvements of the configuration system for crowd-feedback requests in upcoming design cycles.

4.7. Conclusion

While crowd-feedback systems offer a scalable way to collect design feedback, they, however, do not support feedback requesters (e.g., designers) in creating and configuring individual feedback requests. Therefore, in this paper, we present a study on the design of configuration systems that support feedback requesters in individual crowd-feedback request creation. Drawing on existing research on challenges in design evaluation and an interview study with experts, we contribute with four design rationales to support feedback requesters in selecting and configuring feedback features while considering the effects of feedback features on the crowd's perceptions as well as on the feedback quality and quantity. Our results show that feedback requesters appreciated the guidance but leaving too many decisions open made them feel lost in the process. Also, they emphasized the idea of offering templates based on experts' recommendations. Overall, our study contributes design knowledge that can be applied to guide feedback requesters through the decision-making process of creating crowd-feedback requests for the evaluation of software designs. With this, we contribute to making software evaluation more simple, scalable, and efficient and support the development of more human-centered software.

5. Study IV: CrowdSurfer: Seamlessly Integrating Crowd-Feedback Tasks into Everyday Internet Surfing

5.1. Introduction

The continuous evaluation of interactive designs with users is crucial for the acceptance of and user satisfaction with interactive systems (Ives & Olson, 1984; McKeen & Guimaraes, 1997). For example, effective evaluation techniques have been recognized as essential for websites in order to successfully attract customers (Chiou et al., 2010). A typical means for effective evaluation is the collection of user feedback in situ, during the usage of a website. In situ feedback collection using pop-ups or feedback buttons is a powerful way to identify problems, critically reflect on existing features, or collect new additional requirements to increase users' acceptance (Sherief et al., 2014). However, users often perceive such feedback requests as hindering and annoying. In general, the willingness to engage with the feedback request and share meaningful feedback is low (Almaliki et al., 2014). A potential solution to counteract this engagement challenge is the inclusion of paid crowdworkers to gather design feedback, also called crowd feedback. Crowd-feedback systems allow the large-scale collection of feedback via crowdsourcing tasks on platforms like Amazon Mechanical Turk (MTurk) and Prolific (Haug & Maedche, 2021a). Crowd feedback has shown to be a scalable approach for successfully collecting diverse opinions and improving interactive designs (Luther, Tolentino, et al., 2015; Oppenlaender, Tiropanis, & Hosio, 2020; Wauck et al., 2017).

Against these benefits, crowd feedback has major drawbacks: First, crowdworkers lack the actual context of use when providing feedback on interactive systems like websites (Sherief et al., 2014). They are most likely not real users of the respective website and do not really experience it. Potentially, they are not even familiar with the specific context of the website. This might distort their feedback. Offering crowdworkers context has shown to increase their empathy (Ayalon & Toch, 2018) and, in turn, to improve the feedback quality and quantity (Wauck et al., 2017). Context, e.g., in the form of personas has a positive impact on empathy because this helps to recognize and understand the real users'

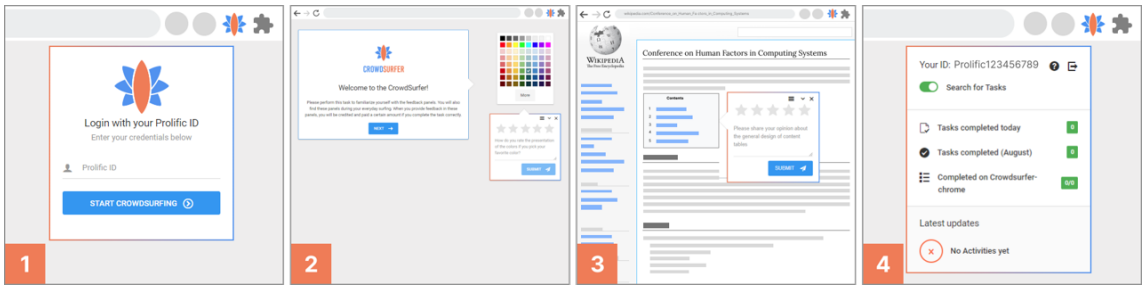


Figure 5.1.: The *CrowdSurfer* extension: 1) Crowdworkers install the *CrowdSurfer* and register with their ProlificID. 2) The *CrowdSurfer* is explained in a demo task. 3) Crowdworkers can solve feedback tasks during everyday internet surfing. 4) Crowdworkers can manage tasks and payments via the *CrowdSurfer* extension.

thoughts and feelings (Detert et al., 2008). Second, if crowd-feedback systems offer an artificial usage context like a scenario or a persona (e.g., Haug and Maedche, 2021b) time and effort for workers to immerse in the artificial usage scenario increase. This additional effort must, of course, also be compensated. A mismatch between the required time and effort for a task and the monetary reward is the main reason for crowdworkers to return, abandon, or reject tasks and is also one of the two causes for crowdworkers' poor hourly wages of around $2 to $5 (Hara, Milland, et al., 2019; Kaplan et al., 2018). Third, crowd-feedback studies on crowdsourcing platforms usually run at a specific point in time. They collect feedback only on a snapshot of the system and do not allow to collect feedback continuously. Due to these drawbacks, in situ feedback is preferable since the feedback providers experience the system and its functionalities in context and real-time (Maalej & Pagano, 2011; Pagano, 2013; Seyff, Ollmann, & Bortenschlager, 2014). However, there is still a lack of knowledge on the differences between in situ feedback and feedback that is collected separately from the actual usage (e.g., in a survey-based crowdworking task), especially regarding feedback quality and quantity. Current research lacks an approach that tackles these three drawbacks of crowd feedback simultaneously. The main focus of research on crowd feedback is still to maximize the feedback quality and quantity. The crowdworker perspective is often neglected (Haug, Benke, & Maedche, 2023). We assume it may be promising to ask crowdworkers for feedback in situ when they are actual users of the system. Simultaneously, this decreases the additional effort of crowdworkers to immerse into feedback tasks, thereby reduces the additional hidden or invisible work for searching, selecting, and accepting the task, and makes the payment fairer. Allowing crowdworkers to solve tasks during their everyday internet surfing might also increase the flexibility of

their working conditions. Ergo, our goal is to integrate feedback tasks into their everyday internet surfing. Thus, we leverage crowdworkers as real users and empower them to work during internet usage. With our field study, we aim to understand how crowdsourced real and in situ user feedback differs from traditional crowd feedback and how crowdworkers perceive the integration of crowdworking tasks into their everyday internet surfing.

Following this objective, in this paper, we present *CrowdSurfer*, an innovative crowd-feedback system in the form of a browser extension that allows crowdworkers to provide website design feedback during their everyday internet surfing in return for a monetary reward. Thereby, we combine the benefits of crowdwork and traditional user feedback. Functionally, the *CrowdSurfer* is connected to a crowdsourcing platform. After installation, crowdworkers can work on existing tasks when visiting respective websites during their everyday internet surfing. Figure 5.1 shows the process of using the *CrowdSurfer* extension as a crowdworker.

We evaluated the *CrowdSurfer* in an experimental field study with 63 crowdworkers following a quantitative and qualitative approach. We assessed the feedback quality and quantity of the design feedback provided as well as the perceptions of the crowdworkers regarding the usability of the *CrowdSurfer*, the feedback process, and its effect on the working conditions of crowdworkers. Our results show that the *CrowdSurfer* was comfortable, simple, and enjoyable for the crowdworkers to use and that they perceived conducting feedback tasks with the *CrowdSurfer* as fairer regarding payment and effort. Although participants stated that they believe the feedback they provided with the *CrowdSurfer* is more real and therefore more relevant, quantitative results showed that the *CrowdSurfer* feedback is less specific, actionable, relevant, and shorter. This discrepancy enlightens an interesting differential between positive effects on the working conditions of crowdworkers and a lower quality of design feedback comments in comparison with traditional feedback tasks. We identified important aspects that demonstrate the utility of the *CrowdSurfer* for requesters despite the reduced feedback quality. Based on our findings we derived recommendations for the future design of crowdsourcing systems integrated into crowdworkers' everyday internet surfing. With our work, we contribute:

- The *CrowdSurfer*, a crowd-feedback system in the form of a Chrome extension for crowdsourcing feedback on websites in return for monetary rewards.

- Findings demonstrating the utility of a browser extension to include feedback tasks in crowdworkers' everyday internet surfing (e.g., with regards to the effort of work, fairness of payment, and flexibility).
- Design recommendations for developing future crowdsourcing systems that integrate tasks into crowdworkers' everyday internet surfing.

5.2. Related Work

In the following, we present related work on crowdsourcing design feedback, casual microtasking and in situ feedback, and crowdworker working types, behaviors, and conditions.

5.2.1. Crowdsourcing Software Design Feedback

User involvement in the continuous evaluation of website designs is crucial. Traditionally, websites are evaluated using methods like usability tests, interviews, or focus groups (Vredenburg et al., 2002). As these methods lack scalability, are costly, and require access to users, crowd feedback has evolved as a complementary approach for collecting large amounts of design feedback. Crowd feedback comes from the visual design domain where feedback is usually provided by peers (Yuan et al., 2016). Various crowd-feedback systems have been suggested to collect quantitative and qualitative design feedback for websites (Haug & Maedche, 2021b; Oppenlaender, Tiropanis, & Hosio, 2020) or mobile apps (Seyff, Ollmann, & Bortenschlager, 2014). Further, such crowd-feedback systems often include numerous design features to enrich the feedback (Haug & Maedche, 2021a). Research has shown that crowd-feedback systems are capable of achieving a feedback quality similar to expert feedback (Luther, Tolentino, et al., 2015). One of the first crowd-feedback systems is *Voyant* (Xu, Huang, et al., 2014). *Voyant* was designed to collect feedback on poster designs by collecting impressions of the crowd and analyzing the adherence to design guidelines. *Voyant* combined the collection of qualitative and quantitative feedback with a marker feature so that feedback providers could draw boxes to highlight a designated area and support their textual feedback (Xu, Huang, et al., 2014).

While some systems similar to *Voyant* are focused on feedback collection during the development process (e.g., Oppenlaender, Kuosmanen, et al., 2021; Schneider et al., 2016; Wauck et al., 2017), others collect feedback after go-live for continuous improvement (e.g., Oppenlaender, Tiropanis, and Hosio, 2020; Seyff, Ollmann, and Bortenschlager, 2014;

Stade et al., 2017). While crowd-feedback systems provide multiple benefits to continuously evaluate the software designs during the development process, they have downsides as well. Often, they only allow feedback collection in dedicated studies and continuous in situ feedback collection is not possible. Furthermore, crowdworkers do not actually use the software and the feedback is provided in an artificial usage context, for example, with a persona as context (Ayalon & Toch, 2019), or a usage scenario to consider when providing feedback (Haug & Maedche, 2021b). Our study extends prior work by offering an innovative way to crowdsource website design feedback. With the *CrowdSurfer* we tackle the mentioned problems by enabling crowdworkers to conduct design feedback tasks during their everyday internet surfing. Additionally, we want to contribute with a better understanding of the impact of the context on feedback quality.

5.2.2. Integration of Feedback Tasks in Internet Surfing

Hahn et al. (2019) invented the term *casual microtasking* to describe the integration of microtasks into other primary activities of workers. In their study, they inserted writing microtasks into the Facebook feed to allow workers to solve microtasks during short breaks. Their results indicate that casual microtasking is a promising approach to leveraging spare micromoments (Hahn et al., 2019). Further studies investigated the role of the context of crowdworkers when accepting tasks. The results of Goncalves, Hosio, et al. (2015) highlight the potential of context to motivate participation in ubiquitous crowdsourcing tasks. They showed that if the crowdsourcing task is located directly next to the physical element on which feedback is collected the participation rate increases. Therefore, situatedness in feedback tasks seems to increase participation rates and engagement. Also, the crowdworker context influences task acceptance and crowdworker preferences (Hettiachchi, Wijenayake, et al., 2020).

The integration of feedback tasks into everyday internet surfing leads to the collection of so-called in situ feedback. In situ feedback is user feedback that is collected while the user is actually using and experiencing the system. A key advantage of in situ feedback is that users do not have to leave the experience to provide feedback which means less interruption to them (Pagano, 2013). There exist dedicated systems to collect in situ feedback. *AppEcho* (Seyff, Ollmann, & Bortenschlager, 2014) is a mobile feedback approach that allows users to provide feedback about their smartphone applications. *iRequire* (Seyff,

Graf, & Maiden, 2010) is a similar system that allows users to provide feedback on their environment, such as a timetable at a bus stop. In the application, they can take a picture and add a textual description of their related requirements. In situ feedback may also be combined with passive logging data as applied by *MyExperience* (Froehlich et al., 2007), a system that captures device usage, user context, and environmental sensing in the background. Additionally, *MyExperience* conducts user experience sampling to collect in situ user feedback.

These studies have demonstrated the feasibility and advantages of capturing in situ feedback. However, existing research has mainly focused on developing mobile applications to capture in situ feedback. In our study, we want to provide a crowd-feedback system for collecting in situ feedback from crowdworkers. To do this, we want to leverage the approach of casual microtasking based on the results of Goncalves, Hosio, et al. (2015), Hahn et al. (2019), and Hettiachchi, Wijenayake, et al. (2020) by providing further insights on how to integrate tasks into crowdworkers' daily life. Thereby, we aim to understand how in situ feedback differs from traditional survey-based feedback.

5.2.3. Working Conditions of Crowdworkers

Crowdwork is a well-researched topic in the field of human-computer interaction. When designing an innovative approach for crowdsourcing tasks, we need to understand the crowdworkers' characteristics, working behavior, and preferences as well as their problems, requirements, and restrictions. This allows for informing the design rationales for our *CrowdSurfer* extension.

Research on crowdworkers' characteristics showed that many of them are multitaskers and mix work and non-work activities (Lascău, Gould, Brumby, & Cox, 2022). This finding is also supported by A. C. Williams et al. (2019) who found out that crowdworkers tend to divide their attention between work and non-work related activities (e.g., watching TV). This may be partly caused by the support tools frequently used by crowdworkers. These tools (e.g., MTurk Suite [1], TurkerView [2]) enable and reinforce task-switching and multitasking behavior. They also promote the fragmentation of crowdworkers' work-life boundaries as they enable a "work-anywhere" attitude (A. C. Williams et al., 2019). To better understand the work practice of crowdworkers, A. C. Williams et al. (2019) also

[1] https://chrome.google.com/webstore/detail/mturk-suite/iglbakfobmoijpbigmlfklckogbefnlf
[2] https://turkerview.com/

investigated the work-life boundaries of crowdworkers. In their study, the majority of participants had a low boundary control, meaning they felt they could not control the timing, frequency, and direction of boundary crossings regarding interruptions to fit their identities (Kossek et al., 2012).

Crowdworking platforms (e.g., MTurk, Prolific, CrowdFlower) differ in the types of users they attract. For example, in the study of Abbas and Gadiraju (2022) 41% of participants on MTurk reported using MTurk as their main source of income, while only 8% of Prolific users reported the same for Prolific (Abbas & Gadiraju, 2022). This is consistent with the results of earlier studies (Berg, 2016) and shows that Prolific workers are potentially more open to casual microtasking as they are not purely focusing on maximizing their financial rewards. Also, the social protection and working conditions of crowdworkers have already been investigated in multiple studies (e.g., Codagnone et al., 2017; Felstiner, 2011; Lascău, Gould, Brumby, and Cox, 2022). Frequently mentioned problems regarding the working conditions are the limited flexibility of crowdworkers (Lascău, Gould, Brumby, & Cox, 2022) and the low payment which is partly caused by invisible work (Hara, Adams, et al., 2018). The term invisible work summarizes unpaid but necessary duties of crowdworkers such as job search, task rejection, task submission, and task information gathering (Hara, Adams, et al., 2018; Saito et al., 2019). To address the mentioned challenges, recent research already proposed extensions for crowdworkers to better manage their tasks, increase transparency, and give crowdworkers a voice (e.g., *TurkScanner* (Saito et al., 2019), *Turkopticon* (Irani & Silberman, 2013), and *Turker Tales* (Kasunic et al., 2019)).

We believe that it is important for our crowd-feedback system to consider crowdworkers' characteristics and enable the conduction of fair crowdworking tasks. This includes supporting crowdworkers in setting boundaries between work and non-work-related activities, counteracting invisible work, and increasing flexibility.

5.3. The Crowd-Feedback System CrowdSurfer

The goal of our design solution is twofold. First, we want to combine the benefits of user feedback, especially the real context, with the scalability of crowdsourcing for design feedback collection. Second, we want to improve the working conditions for crowdworkers and provide them with a fair and flexible way of working. Therefore, we decided to design

a crowd-feedback system in the form of a browser extension that enables crowdworkers to gain monetary rewards during their everyday internet surfing.

Before developing a full-functioning browser extension, we first developed an initial prototype, which we then discussed in exploratory interviews with five crowdworkers (four female, one male). The participants were recruited on Prolific and were on average 34.80 years old (SD = 10.76) with one to seven years of crowdworking experience. These interviews helped us to elaborate our design rationale and understand how distinctive features need to be implemented in the final *CrowdSurfer* system.

5.3.1. Design Rationale

With respect to the goals of our study, the design of our crowd-feedback system follows three fundamental design rationales:

1. *Seamless integration in everyday internet surfing.* The main goal of our crowd-feedback system is to allow seamless integration of design feedback tasks in crowdworkers' everyday internet surfing. Crowdworkers shall not be distracted from their primary tasks but still notice the availability of feedback tasks. To achieve high adoption, it is crucial that users do not get annoyed by feedback requests.

2. *Control for crowdworkers.* As our feedback extension impacts crowdworkers during their everyday internet surfing, they need to be in control over the system in general and the tasks in particular. They also need to be able to control their boundaries between non-work and work activities.

3. *Feedback value.* The system needs to generate high-quality feedback to present value to feedback requesters. The system shall be able to collect different types of feedback to address requesters' needs.

5.3.2. System Design

In this chapter, we will present the final design of the *CrowdSurfer* and the implemented features. We explain our design decisions by referring to related work or to our exploratory interviews. To allow for seamless integration into crowdworkers everyday internet surfing, we decided to implement the crowd-feedback system as a Chrome extension. This Chrome extension displays feedback tasks as pop-ups on the respective websites. In the following,

we describe the design and features of the *CrowdSurfer* according to three steps: (1) Download and setup, (2) providing feedback, and (3) managing tasks.

5.3.2.1. Download and Setup

The *CrowdSurfer* can be installed via the Chrome web store. Crowdworkers can log in by entering their crowdsourcing platform ID (here Prolific) (see Figure 5.2 (top)). This is required so that their task submissions can be matched to and paid via their crowdwork account. Then, crowdworkers need to conduct a demo task to learn about the features of the *CrowdSurfer* (see Figure 5.2 (bottom)).

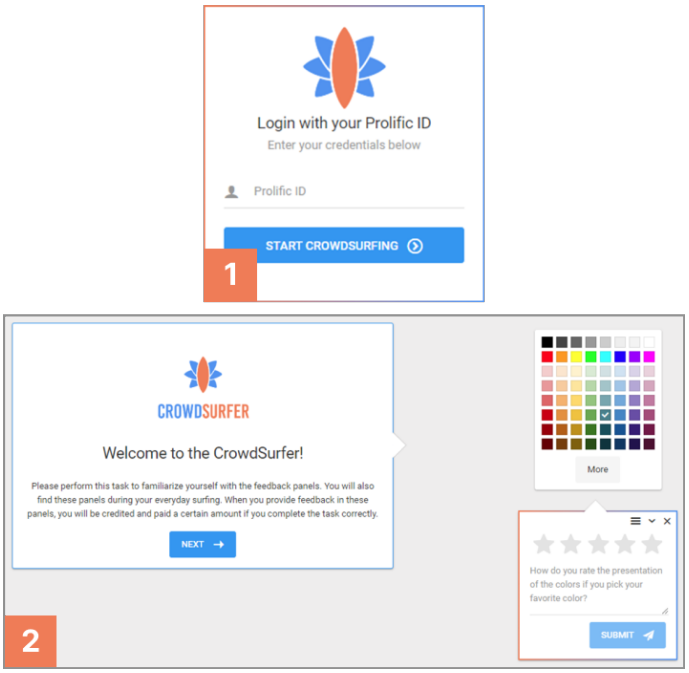


Figure 5.2.: Setup of the CrowdSurfer: 1) Login screen, 2) demo task to explain features

5.3.2.2. Providing Feedback

After the setup and the demo task are completed, the *CrowdSurfer* will display feedback tasks as pop-ups on selected websites. A screenshot of such a feedback pop-up is displayed in Figure 5.3 (left). Each feedback pop-up is attached to a website element (1). The evaluation of website elements instead of a whole website, in general, allows feedback requesters to get more specific and structured feedback from users. Additionally, crowdworkers feel that providing feedback is easier when they can focus it on a specific element (Haug, Benke, & Maedche, 2023). In our interviews, participants initially complained that "*it's not clear if I have to rate the whole area or just a part*" (I3). Consequently,

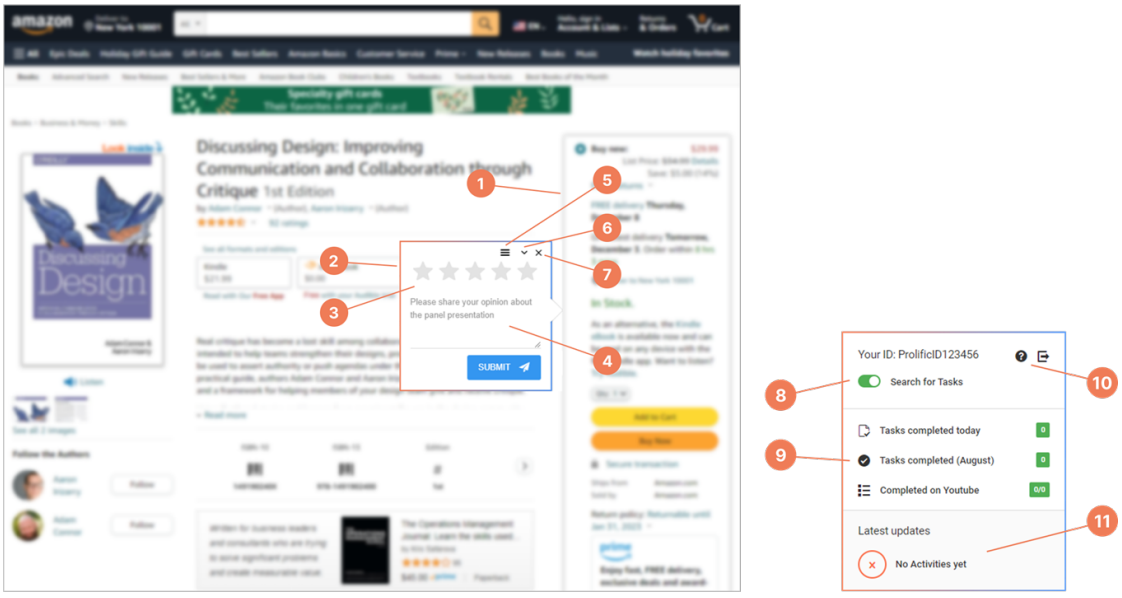


Figure 5.3.: Feedback pop-up on blurred Amazon website (left) and *CrowdSurfer* panel (right): 1) Element on which the feedback is collected, 2) feedback request pop-up, 3) star rating, 4) feedback text field with a question, 5) menu icon to see background information and set a reminder, 6) minimize icon, 7) reject icon, 8) toggle button to turn the *CrowdSurfer* on and off, 9) information on task rewards, 10) support icon to redo the demo task, and 11) overview of recently submitted tasks.

the design of the feedback pop-ups was refined so that they clearly highlight a website element. When collecting crowd feedback from real users, usually both qualitative and quantitative feedback is collected (Haug & Maedche, 2021a). Both feedback types have their advantages and disadvantages. Feedback providers often prefer multiple choice or ratings as it is simpler and faster (Almaliki et al., 2014), but qualitative text feedback of course contains more information. Therefore, our feedback extension is able to collect both types of feedback. The quantitative feedback is collected in form of a star rating (3) and the qualitative feedback as an answer to a question about the respective element (4). Each feedback pop-up has a menu (5) that allows crowdworkers to access task information of a task, e.g., the payment, the requester name, or contact information. Thereby, we want to counteract the information imbalance between feedback requesters and crowdworkers. While feedback requesters can access a lot of information about the crowdworkers, such as qualification, location, or experience, crowdworkers usually can only access limited information such as creation date and reward amount (Kaplan et al., 2018). The menu allows workers to set a reminder for the task, in case they want to postpone it. Postponing of tasks would for example be helpful if *"you find a task and you realize that it's going to take*

longer than what you thought it would and you'd like to go back to it and finish it later" (I1). This feature also provides workers more control over how they want to do their work. Next to the menu is a button to minimize the task (6) to allow a seamless integration into the website. The pop-up could hide important elements of the website and crowdworkers shall not be forced to complete or reject the task just to be able to see the whole website. Finally, the cross icon (7) allows crowdworkers to reject tasks if they are not interested in solving them. After submitting a task, crowdworkers can see how many tasks on the website they solved and how many are remaining (e.g., "1/2 tasks completed").

5.3.2.3. Managing Tasks

When clicking on the icon of the *CrowdSurfer* in the list of extensions, crowdworkers can access a pop-up to manage their tasks and the *CrowdSurfer* extension. This pop-up is displayed in Figure 5.3 (right). To give crowdworkers control over the extension and their work-life boundaries, they can turn it off when they do not want to see any feedback tasks (8). In our interviews, workers stated that there are situations in which they do not want to be interrupted by such an extension (*"If I'm in an interview and then I keep on being distracted by this thing that keeps popping up, then it won't work. So I would like to be able to turn it off"* (I1)). This feature also addresses crowdworkers' concerns regarding their data privacy as they were worried about the extension always tracking their online behavior and data. Next, the pop-up should contain an overview of the number of completed tasks so that crowdworkers "*see whether or not it's worth your time*" (I1). As P4 stated that they "*prefer it showing more of their earnings as opposed to how many tasks you sold because the earnings can help you dictate how much you're going to earn in total, sort of a certain target*", we do not show the number of tasks but the total reward for the day and the month (9). Additionally, the panel shows how many tasks are open on the current website, so that crowdworkers do not start searching for tasks when there are none. In case crowdworkers want to redo the demo task to learn about the CrowdSurfer's functionalities again, they can do this by clicking on the question mark icon (10). Finally, crowdworkers requested in our interviews to somehow be able to see what the last task was that they submitted because "*in an instance where you had a network issue [...] that latest update is going to be helpful for you to go in and see if you really have completed the task*" (I5). We show this information in the latest updates (11).

5.4. Evaluation Study

To analyze the effects of the *CrowdSurfer* we conducted a quantitative and qualitative field study with crowdworkers on Prolific. The goal of this field study was to understand the benefits of *CrowdSurfer* feedback compared to traditional survey-based design feedback in terms of feedback quality and quantity. Further, we aimed to understand the impact of the *CrowdSurfer* on the working conditions of crowdworkers such as effort of tasks, fairness of payment, and flexibility.

5.4.1. Procedure

For the evaluation, we decided to implement in total 13 *CrowdSurfer* tasks on eight of the most frequently used websites (*YouTube*, *Amazon*, *Twitter*, *Wikipedia*, *eBay*, *CNN*, *Weather.com*, and *Reddit*) to ensure that participants will visit the websites coincidentally. Participants had seven days to use the *CrowdSurfer* and provide feedback on the respective websites. For each star rating, they received £0.03, and for each text feedback £0.12. After seven days, participants were notified that they could now participate in a post-task questionnaire on Prolific. This questionnaire also offered them the option to schedule a 20-minute interview with us in return for a £4 bonus payment. For the baseline treatment, we developed a simple feedback task with a questionnaire that showed links and screenshots of websites and asked for feedback on specific elements (see Figure 5.4). The payment per feedback was the same as in the *CrowdSurfer* treatment. Afterward, they also received the same post-task questionnaire. In the post-task questionnaires, we included several attention checks. The websites and tasks were in both treatments the same and participants could in both treatments freely choose if they wanted to provide feedback or not. We performed twelve semi-structured qualitative interviews with participants of the *CrowdSurfer* treatment to understand how crowdworkers perceived the extension and to interpret the quantitative results. We focused in our interviews on three aspects: the usability of the *CrowdSurfer*, the crowdworkers' feedback process, and the impact of the *CrowdSurfer* on the working conditions of crowdworkers including their motivation. The study was approved by the German Association for Experimental Economic Research (GfeW).

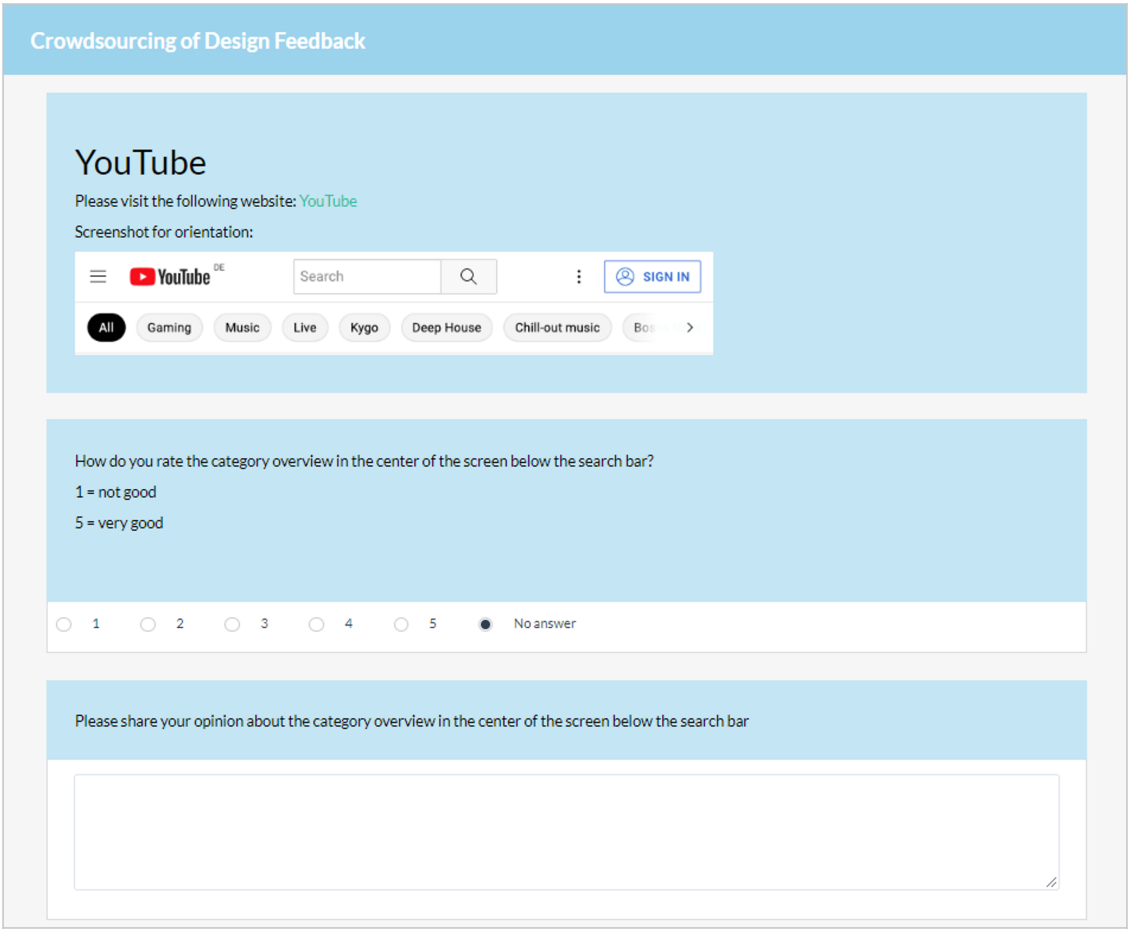


Figure 5.4.: The design feedback survey that we used in our baseline condition.

5.4.2. Participants

We recruited 80 participants (30 baseline, 50 CrowdSurfer). We filtered for English native speakers to achieve a comparable feedback quality between the participants. In the *CrowdSurfer* condition, nine participants downloaded the *CrowdSurfer* but never submitted a task which resulted in 71 participants. Of these 71 participants, eight participants failed the attention checks in the questionnaire. So, we ended up with 63 final participants in our sample (29 baseline, 34 CrowdSurfer). 36.5% of the participants were female (63.5% male) and their average age was 35.8 years (SD = 12.92). Overall, the participants reported a medium experience in providing design feedback on a seven-point Likert scale (M = 3.09; SD = 1.21). They were randomly assigned to the experimental conditions.

5.4.3. Data Collection & Analysis

Quantitative data analysis. We gathered data in two ways. First, via the post-task questionnaire, we collected data on the perceived task duration, the perceived fairness of pay-

ment, and the perceived flexibility. For the perceived fairness of payment, we reused the items of Schulze et al. (2012) that were previously used to measure the fairness in pay (Alpar & Osterbrink, 2018). For the perceived flexibility we adapted the items of Kokoç (2019) and Richman et al. (2008).

Second, we analyzed crowdworkers' interaction with the *CrowdSurfer*. We tracked how many tasks were shown and how they interacted with them. We logged when they interacted with one of the features of the *CrowdSurfer* (reminder, minimize task, reject task, show task information, turn the search for tasks on/off). Additionally, we collected and analyzed the feedback that crowdworkers gave in both treatments.

Feedback quality evaluation

To analyze the quality of the collected feedback comments, we conducted a separate crowdsourcing task in which UI design experts assessed the quality of the design feedback comments in six dimensions. For this task, we again used Prolific where we were able to filter for crowdworkers with UI design skills by using the respective filter. Thereby, we recruited 103 crowdworkers with experience in UI design (M = 4.51, SD = 1.53, self-assessed on a seven-point Likert scale) for the quality assessment.

To provide feedback evaluators context, we sorted the feedback comments according to the website and element they belong to. For each website, we created separate tasks that showed each crowdworker on which element the comment was provided and presented them with up to 20 feedback comments to assess. Each feedback comment of both treatments was analyzed by three participants on the following dimensions: specificity, explanatory, actionable, positivity, relevance, and overall feedback quality. The dimensions were adopted from the study of Oppenlaender, Kuosmanen, et al. (2021). The feedback quality value for each construct was assessed by taking the average from the distinct ratings of the three individual crowdworkers.

Qualitative data analysis

The interviews were transcribed and analyzed by two of the authors. We analyzed the feedback through a deductive thematic analysis following Braun and Clarke (2006) based on the main topics of our interviews: *CrowdSurfer* usability, the feedback process, and the working conditions of crowdworkers. After the deductive analysis, we inductively refined

the coding scheme. Finally, all interviews were coded by two authors. Disagreements were discussed until a consensus was found.

5.5. Results

To investigate the effect of the *CrowdSurfer* on the design feedback and the crowdworker experience we conducted a three-folded analysis. First, we present the usage behavior of crowdworkers with the *CrowdSurfer* based on the log data. Second, we present results on crowdworkers' perceptions regarding the perceived time they spent working on tasks, the fairness of payment, and their flexibility. Further, we present results on the design feedback quality and quantity. In the third part, we present the themes that resulted from our qualitative interviews.

5.5.1. CrowdSurfer Usage Behavior

In this section, we describe the crowdworkers' behavioral interaction with the *CrowdSurfer* based on our log data. The results are displayed in Table 5.1. In total, participants solved 240 tasks of the *CrowdSurfer* within the experimental period of seven days. 15 of these tasks only contained a star rating. While 50 crowdworkers installed the *CrowdSurfer*, only 41 provided feedback at least one time. The majority of crowdworkers provided feedback between two and seven times.

The most frequently used feature after the submit button was the toggle button which turns the *CrowdSurfer* off. In this mode, no feedback pop-ups are displayed. Crowdworkers solved more than half of the tasks in the first two days after installing the *CrowdSurfer* (see Figure 5.5). Of the eight websites on which tasks were available, crowdworkers provided the most feedback on YouTube, followed by Amazon and Weather.com. Although crowdworkers did not complete every task the first time it was displayed, for 87.59% of the displayed tasks crowdworkers submitted feedback, eventually. On average, crowdworkers submitted tasks 69.23 seconds (SD = 49.20 seconds) after entering the website.

5.5.2. Working Conditions & Feedback Quality

5.5.2.1. Working Conditions of Crowdworkers

To assess the crowdworkers' perceptions, we analyzed the responses to questionnaire items. To assure the internal consistency of latent constructs, we assessed outer factor loadings

Table 5.1.: *CrowdSurfer* feature usage by crowdworkers.

Feature	Used ... times	Used by ... crowdworkers	Average usage per worker
Task information	4	3	1.33
Reminder	7	3	2.33
Minimize	10	8	1.25
On/off task search	96	36	2.67
Feedback submit	240	41	5.85

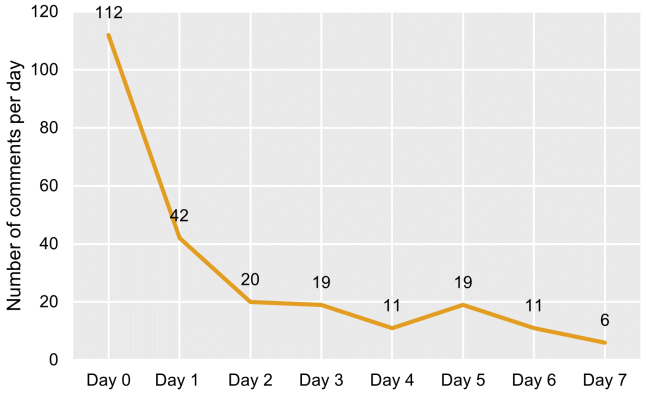


Figure 5.5.: Submitted tasks per day over the period of seven days.

and Cronbach's alpha with a cutoff at 0.7 and 0.6 (Hair et al., 2014; van Griethuijsen et al., 2015). Afterward, scales were mean-scored. To assess the effect of the experimental treatment conditions (baseline vs. *CrowdSurfer*), we conducted an analysis of variance (ANOVA) for each variable and the feedback quality and quantity assessments as dependent variables. The results show no significant results for work flexibility, but a significant effect for fairness of payment ($F(1,61) = 5.76$, $p < 0.05$) between the *CrowdSurfer* and the baseline treatment. For the perceived time invested to complete the task, we find a significant effect ($F(1,61) = 4.02$, $p < 0.05$). Detailed information regarding descriptive statistics is presented in Table 5.2. To complement the quantitative analysis, we present boxplots of the perceptive measures in Figure 5.6.

Table 5.2.: Descriptive statistics of perceptive measures over the two treatment conditions.

Dependent variable	Details	Baseline (n = 29)	CrowdSurfer (n = 34)	Analysis results
Work flexibility	Mean (SD)	5.335 (0.987)	5.273 (0.824)	Not significant, $F(1,61) = 0.07$, $p = 7.87$
Fairness of payment	Mean (SD)	4.931 (1.665)	5.735 (0.946)	Significant, $F(1,61) = 5.76$, $p < 0.05$
Perceived task completion time	Mean (SD)	11.241 (6.098)	8.029 (6.530)	Significant, $F(1,61) = 4.02$, $p < 0.05$

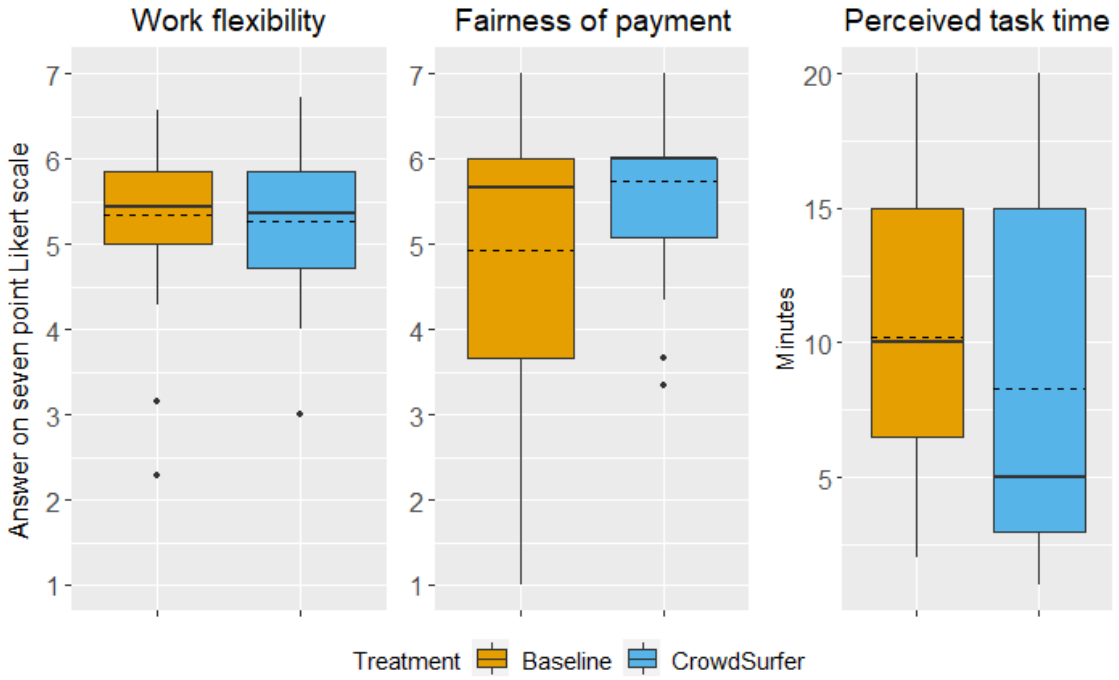


Figure 5.6.: Boxplots of perceptions of work flexibility, fairness of payment, and time for task completion (the dotted line represents mean value).

5.5.2.2. Design Feedback Quality

For the six design feedback quality dimensions, we performed ANOVAs to test the effect of the treatment on the dependent variables for the feedback comments. For almost all variables we see a positive main effect between the baseline and the treatment group with higher values for the baseline condition (see Table 5.3 for detailed results of the ANOVA tests and Figure 5.7 for the boxplots). Only for positivity, we see a higher level in the *CrowdSurfer* condition and a not significant main effect ($F(1,563) = 0.058$, $p = 0.81$). Further, we analyzed the difference in the length of the feedback comments provided by the participants. To do so, we analyzed the number of characters per comment. The results of the ANOVA showed a significant main effect ($F(1,563) = 9.26$, $p = 0.01$) with longer comments in the baseline condition.

5.5.3. CrowdSurfer Experience

We analyzed and coded the interviews to understand how crowdworkers perceived the *CrowdSurfer* for conducting feedback tasks. We derived 20 themes that describe crowdworkers' positive and negative experiences with the *CrowdSurfer*. Overall, all participants liked the concept of the *CrowdSurfer*. They found conducting tasks with the *CrowdSurfer*

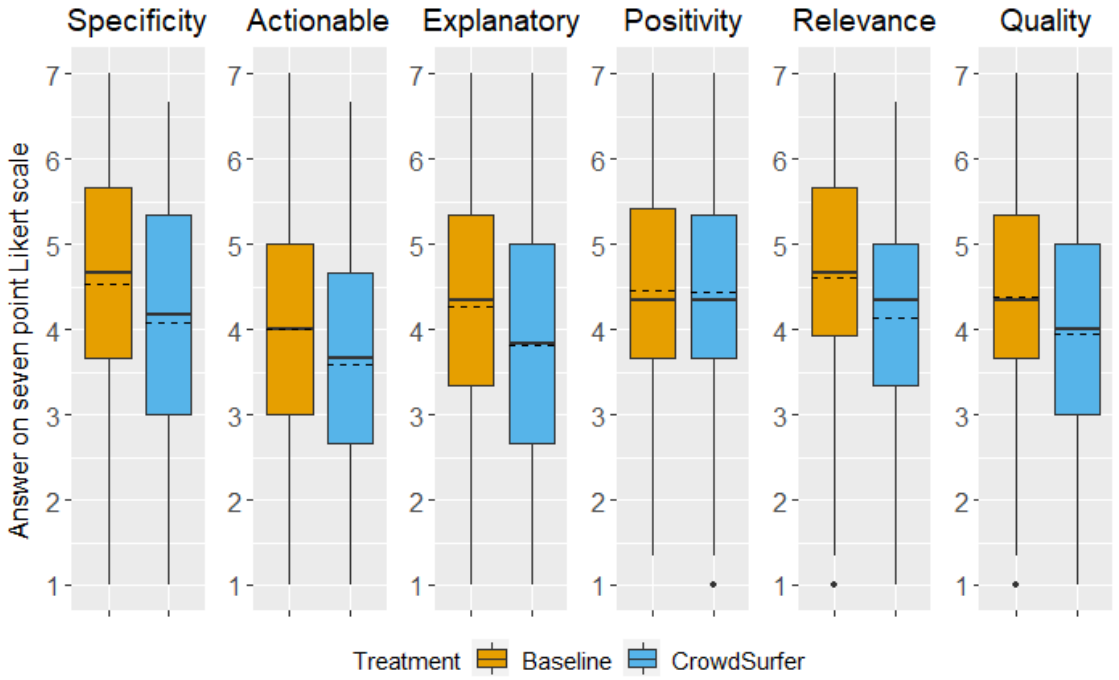


Figure 5.7.: Boxplots of design feedback quality dimensions (based on feedback comment level, dotted line represents mean value).

Table 5.3.: Statistics of design feedback quality dimensions over the two treatment conditions, aggregated on comment level.

Dependent variable	Details	Baseline (comment, n = 340)	CrowdSurfer (comment, n = 225)	Result
Specificity	Mean (SD)	4.529 (1.367)	4.078 (1.441)	Significant, $F(1,563) = 14.16$, $p < 0.01$
Actionable	Mean (SD)	4.010 (1.268)	3.593 (1.282)	Significant, $F(1,563) = 14.29$, $p < 0.01$
Explanatory	Mean (SD)	4.273 (1.501)	3.803 (1.516)	Significant, $F(1,563) = 13.15$, $p < 0.01$
Positivity	Mean (SD)	4.455 (1.261)	4.430 (1.221)	Not significant, $F(1,563) = 0.058$, $p = 0.81$
Relevance	Mean (SD)	4.609 (1.239)	4.133 (1.273)	Significant, $F(1,563) = 19.61$, $p < 0.01$
Overall Quality	Mean (SD)	4.381 (1.341)	3.948 (1.355)	Significant, $F(1,563) = 14.01$, $p < 0.01$
Comment Length	Mean (SD)	141.674 (116.017)	112.009 (109.458)	Significant, $F(1,563) = 9.26$, $p < 0.01$

easy and fun, and experienced the interaction as comfortable. However, they still experienced issues and raised concerns, especially regarding the clarity of the tasks and the security of their personal data when installing an extension. An overview of the positive and negative aspects of the *CrowdSurfer* based on our interviews is shown in Table 5.4.

Table 5.4.: Overview of positive and negative aspects of the *CrowdSurfer* derived from the qualitative interviews.

	Positive aspects	Negative aspects
CrowdSurfer usability	- simple to use - fun to do tasks - comfortable and organized - the combination of quantitative and qualitative feedback - more personal feedback requests	- requests are not specific enough - UI design could be improved - security and data privacy concerns
Feedback process	- seamless integration in normal internet surfing - more in the usage context - transparent regarding what and when data is collected	- interruption of other tasks - not all tasks will be found - could get repetitive
Working conditions	- no pressure in terms of if, when and how to do tasks - payment higher than expected - less effort for background work - no feeling of scarcity - multiple incentives besides monetary reward	- some participants searched for tasks - no trust that feedback will be used

5.5.3.1. CrowdSurfer Usability

In the first part of our interviews, we asked participants how they perceived the interaction with the *CrowdSurfer* in general. Ten participants mentioned that they appreciated that the *CrowdSurfer* was so easy to use and perceived the interaction as very comfortable (*"It was simple, I mean anyone could do it. There's nothing technical about this. It just asks a question and you give your answer"* (P4)). Three participants stated *"it was fun"* (P5) to provide feedback with the *CrowdSurfer*.

They perceived the *CrowdSurfer* as a non-intrusive, transparent, and seamless way to provide feedback (*"I knew it was there in the background. That was one thing that was good. It wasn't hiding."* (P4)). The feedback requests felt for three participants very personal. P5 explains *"I think it also kind of feels a little more personalized in a way because it's not just like this survey form that you fill out and everybody fills out the same thing. Like when it pops up on your screen, while you're doing the browsing, it kind of feels more intimate [...]"*. On the negative side, five participants reported that they were worried about the security and their data privacy when installing the *CrowdSurfer*: *"I think the only issue is when people might think 'Well, hang on a minute, can I trust this to be on all the time, or should I turn it off when I'm banking or checking personal information?'"* (P3). However, overall they perceived the *CrowdSurfer* as trustful enough to still decide to install the extension. Four times participants mentioned that they would like to have

more background information about why and by whom the feedback is collected to better target their feedback because *"what was there was quite basic"* (P4) and "*it can be a little bit ambiguous because you could always be rating different facets of whatever X is"* (P12). Furthermore, they perceived the feedback requests as *"vague"* (P6). While it was seen as positive that the feedback pop-ups blend in with the original website, two participants were worried that they would miss feedback requests because the pop-ups do not stand out enough. One participant felt like the UI design could be *"a little more advanced"* (P5).

CrowdSurfer vs. traditional crowdsourcing tasks

Seven participants explained that the main difference between providing design feedback via the *CrowdSurfer* and doing it in a survey is that they felt more in the context of use: *"I think that I like this better than just filling out a normal survey because when they're asking questions it's about what I'm seeing right there in front of my eyes, so I don't have to rely on my memory of the experience [...]. I'm in the experience, I can read it, I can say what I think."* (P5). The feedback situation is *"more direct"* (P2), and questions are asked *"at the relevant time"* (P1). The crowdworkers are not "*in the mindset of being paid to go through a website and break it down and try to find things wrong with it"* (P11) and are therefore able to *"give a more authentic answer"* (P11). They even spent less time thinking about their feedback, which made them feel like their feedback gets more valuable. Three participants perceived the *CrowdSurfer* as being more comfortable than filling out a survey because although they thought the pop-ups were surprising and random, the tasks felt more predictable *"in terms of when and how many tasks you might do"* (P12). Additionally, answering the *CrowdSurfer* questions felt like less effort than doing the same in a survey.

CrowdSurfer features

Regarding the features, four participants in our interviews mentioned that they used the on-off toggle button. They used it either to turn off the *CrowdSurfer* when they did not want to be traced or interrupted or to refresh the available tasks. They thought when turning the *CrowdSurfer* on again new tasks might pop up *("I just wanted to see, if it's gonna be giving me tasks if I switch it on and off"* (P6)). Five participants used the overview to check their rewards or last tasks. Two participants liked that they could see how many tasks they already found and solved on the current website so that they knew

when they could stop looking out for tasks. P4 explained: *"I saw that I'd done two out of two tasks, so I knew I didn't have to go around and browse on Amazon anymore. It was done and dusted. [...] You know exactly where you stood."* The reminder and minimize functionalities were not used very frequently. Nevertheless, four participants still found that these might be useful for situations in which they *"didn't have the time or [...] didn't have the mind to take a pause [...]*" (P9).

Ideas for improvement

Addressing the perceptions that the feedback requests and the overview panel could have been more detailed, participants recommended features *"to compare with other people who use it"* (P4), *"a little history of what was going on"* (P4), or *"an up to date list of all of the sites that you could sign in"* (P2). Regarding the list of available tasks, P2 argued *"you wouldn't have the problems of random sort of winning the lottery by getting a website where there is a question. If people were expecting to actually earn the money regularly doing this task, I think they'd have to have the structure of a list rather than the frustration of just sort of wandering around and hoping that one of the sites was on the list [...] I would see that as a waste of my time, and I'm not sure I would take it seriously. I think I'd go and do something else"*. Four participants also asked for more interactive ways to provide feedback (*"Either use phrases that people can choose from or numbers, or they can drag their mouse from one point to the other just kind of engage people in different ways, you can more interact. If people prefer one type of feedback over the other then at least have that variety."* (P5)). As two participants stated that they sometimes accidentally submitted their feedback too early they asked for a way to call back the feedback. P1 stated that it *"would be useful to have that as a feature where you can go 'hang on, I forgot to say this'."* Finally, P4 suggested making the *CrowdSurfer* more intelligent so that it recognizes when the user is willing to provide feedback.

5.5.3.2. Feedback Process

Mainly there are two different users types: Either, they want to solve the tasks as quickly as possible and actively search for the tasks (*"I don't think I would simply wait for something to randomly appear. If I've promised to do a task I like to have a list of what the expectations are and go and do them"* (P2)), or they waited for tasks to pop up during their everyday internet surfing (*"I didn't get to the point where I had to search for a task"*

(P7)). P3 stated: *"I didn't modify my behavior because the crowdsurfing app was there. I just did my normal thing"*. The group of participants who waited for tasks to pop up was much larger than the other one. Five participants declared that they provided feedback usually directly after they saw the pop-up: *"It popped up and just straight away I put in the information"* (P4). Thereby, they mainly shared their quick and immediate reaction to the question because they believed *"often a quick response is the right one"* (P3). Four participants stated that they always provided feedback when they noticed a feedback request. For example, P6 disclosed: *"I didn't decide. I just had to do it for each task that I was given. Like there's none that I saw, and I was like 'No, I'm not doing this one.'."* Situations in which they did not provide feedback were when they were *"really in a hurry"* (P3) or on websites, they *"consider to be unpleasant"* (P2). One drawback for five participants was that they could get interrupted by feedback requests when doing important primary tasks. P11 explained her concerns as follows: *"I could see it get a little bit frustrating because I'm here on Amazon because I need to buy something and Amazon is distracting enough to have another thing pop up and inhibit my shopping process."*

5.5.3.3. Working Conditions

Fairness of Payment

Although multiple participants mentioned that they actively searched for tasks, four participants felt the invisible work to be less than in traditional crowdsourcing tasks (*"With this one, it was easier because [...] the only thing I had to do was to review. The [demographic] background information is already there"* (P6)). Three participants experienced that solving tasks with the *CrowdSurfer* requires less effort and time for preparation before the actual task because *"there's less background work that needs to be done"* (P1). One participant was even surprised about how much money she made when checking the rewards for the first time.

Flexibility

Overall, they perceived the *CrowdSurfer* setup as very flexible. Six participants liked that it felt not pressured (*"It was super chilled. There was no pressure in terms of time and I could do it whenever I wanted [...]. So it was super comfortable, better than the Prolific site"* (P6)) or if to work at all as *"there is no penalty for not giving feedback"* (P11). In

contrast to doing tasks on Prolific, they did not have to be online at a specific time when new tasks are published. This reduced *"that feeling of scarcity around it"* (P11). They liked that *"[...] there is a steady supply of work that could be done"* (P12). However, some workers also stated that the task setup did not have an impact on their flexibility in doing tasks. They also did not see a significant impact on their work-life balance, as P3 explained: *"I literally just did my normal day, nothing to do with work/personal life balance, nothing like that was affected by it at all"*. This was mainly because they could turn it off when they did not want to be interrupted: *"I think I would like both options to be available to me and that I'd be able to choose, and for that choice to be inconsistent. So like one day if I feel like I want to browse and I want to also be able to make some money on the side, then I'd be able to toggle it on to activate it [...] Some other days I might feel like [...] I don't want anyone to be asking me things [...], so I'd be able to have it off, and then it wouldn't pop up. But I think both options can be very useful"* (P5).

Motivation

The main reason for providing feedback was the monetary reward (*"Mostly it was for money"* (P6)) as mentioned by seven participants. However, participants also liked that they were able to share their opinions (four participants), help us with our study (three participants), be able to improve the websites (six participants), or were just curious (two participants). One participant also liked that she now *"actually understood what it takes to write a review"* (P6). Participants felt like they could make an impact with their feedback by contributing to a bigger picture. However, some workers did not care about the impact. Although they felt quite competent to provide meaningful design feedback, especially for websites they visit frequently, two participants mentioned that they would be able to provide better feedback if more background information on the task was provided. They had questions like *"What is she specifically looking for here? [...] What will he use the feedback for? Why is it important to be concerned about the colors?"* (P6). Also, they felt that the feedback pop-up did not encourage them to be reflective, as P3 phrased: *"It didn't encourage me to be reflective. It kind of encouraged me to give a quick response"*. It helped three participants that they already had an opinion for the websites that they were familiar and they *"just answered the question based on [their] experiences"* (P3). Additionally, the tasks were so easy that everyone could do them. Detrimental for the motivation of two participants was that they *"don't really trust companies that ask for feedback in general*

because they never act upon it" (P3).

5.6. Discussion

The majority of the crowdworkers using the *CrowdSurfer* perceived the provision of feedback as more comfortable, simple, fun, and personal than in a traditional design feedback survey. Further, these crowdworkers perceived the payment as fairer and spent less time on the task. On the other side, the feedback collected with the *CrowdSurfer* was less specific, actionable, and relevant, contained fewer explanations, and was of lower quality. Our participants mentioned potential reasons for the reduced feedback quality such as the divergence between a primary and a secondary task in the *CrowdSurfer* treatment. In the following, we discuss three essential theoretical and practical implications of our study and present design recommendations for the design of crowd-feedback systems for everyday internet surfing.

5.6.1. Integrating Crowdsourcing Tasks in Crowdworkers' Everyday Internet Surfing Leads to Less Effort

Over the years, many researchers have argued for higher payments of crowdworkers (Hara, Adams, et al., 2018), especially considering the balance of effort and payment (Kaplan et al., 2018). Further, they advocate for more flexible working conditions (Lascău, Gould, Brumby, & Cox, 2022; Whiting et al., 2019). In our study, one main effect of the *CrowdSurfer* was its positive impact on these working conditions such as the fairness of payment, the time spent solving tasks, and work flexibility. Participants stated in the interviews that the seamless integration made it easier and created less effort for them to provide feedback compared to traditional tasks on crowdsourcing platforms. Our quantitative survey results reveal that crowdworkers actually spent less time working on feedback tasks and perceived the payment as fairer than those in the baseline treatment. These two results make sense. Perceiving tasks to be of less effort makes the payment seem higher and thereby fairer. However, our results are twofold regarding the work effort. Although the time for searching for tasks was significantly lower than in the baseline treatment and most participants reported that they had no invisible work, a few workers reported that they actively searched for tasks. However, searching for tasks was not possible in the baseline. Thinking about a long-term scenario in which the *CrowdSurfer* continuously offers new

tasks to crowdworkers, searching for tasks would become even more counterproductive, especially when the tasks are published on less popular websites.

Crowdworkers did not feel like the *CrowdSurfer* had a negative impact on their flexibility or their work-life balance. The main reason for this was the functionality to turn it on or off whenever they like. The on/off feature not only helped them to ensure that they were not interrupted when working or doing other important tasks but also allowed them to guarantee that their interaction and data were not tracked when surfing privately on the internet (cf. for internet banking). Consequently, we believe the on/off feature is a core element that made crowdworkers feel flexible and comfortable when working with the *CrowdSurfer*. To sum it up, the *CrowdSurfer* not only leads to less effort for crowdworkers but also offers a higher hourly wage to crowdworkers and allows them to be more flexible when working on tasks.

5.6.2. The Quality of In Situ Feedback is Lower than in Dedicated Surveys but the Feedback is More Real

Our data shows that the feedback quality is in most dimensions worse when collecting feedback via the *CrowdSurfer*. According to our interviews, crowdworkers believe that their feedback is still more valuable and real when they provide it in situ and for websites they frequently use. But why is the feedback quality worse? Why does the real usage scenario not lead to more relevant and actionable feedback? Is our proposed approach still a successful model for crowdsourcing design feedback?

First, a potential explanation for the reduced feedback quality is the shorter length of the feedback comments that contain fewer details. One reason for shorter comments could be the different sizes of the text fields for feedback comments. In the baseline, the text field was bigger, which might have led crowdworkers to think they needed to write more. Second, in our interviews, participants stated that they mainly shared their quick reactions to the question. This is consistent with the log data that showed that on average the feedback tasks were submitted about one minute after the participants entered the website. They unconsciously provided quick feedback which came directly ”from the heart”. The main difference between the two treatments was that in the baseline providing feedback was the primary task on which participants were focused, while when using the *CrowdSurfer* providing feedback was a secondary task, and participants potentially focused on another

primary task. We assume that the *CrowdSurfer* treatment group spent overall less time and effort on the feedback provision process as the initial effort was close to zero. Compared to the baseline treatment, they did not have to spend time entering a website, getting familiar with the element on which the feedback is collected, and forming an opinion. They only had to document their thoughts and perceptions. As we learned in our interviews, participants were not only motivated by monetary compensation. Consequently, they might care less about receiving the monetary reward and, in turn, put less focus on their feedback quality, and more on the feedback honesty.

Third, one potential side-effect of the *CrowdSurfer* might be that it favors a special character of crowdworkers. Crowdworkers who are willing to install an extension and are open to an innovative task form might have special approaches to crowdwork. They are more flexible and might be less focused on maximizing their financial outcomes. We assume that these workers do not use Prolific as their primary source of income, but are rather *part-time* crowdworkers.

Finally, our interviewees mentioned that *CrowdSurfer* users did not feel like being paid for finding problems on a website like in the baseline survey. The focus on the problem-finding task itself might have created the perceived urgency to report design flaws. This could be explained by a social desirability bias (Grimm, 2010). Participants know that they are explicitly recruited to highlight design errors. In consequence, they come up with issues, even though these might not represent actual impressions. In the *CrowdSurfer*, the collected feedback on issues is more subconscious and, thus, closer to participants' perceptions. One might say the feedback is more real and less biased and therefore superior to the traditional survey design feedback. Drawing on Goncalves, Ferreira, et al. (2013), exploring motivational factors besides money could be a useful approach to increasing the feedback quality and understanding what makes crowdworkers report design issues besides the monetary reward. Our results also align with related studies on integrating secondary crowdsourcing tasks in primary tasks (Hahn et al., 2019). Although we followed a different motivation, we also come to the conclusion that integrating feedback tasks into crowdworkers' everyday surfing is overall a successful way to accomplish meaningful design feedback. Consequently, we argue that the *CrowdSurfer* is a valuable approach to collecting honest and unbiased design feedback in comparison to traditional surveys.

5.6.3. Crowdworker Archetypes: Seamless Integration vs. Waiting for Tasks

The *CrowdSurfer* was intended and designed for crowdworkers who want to solve tasks and earn money while doing other primary tasks and therefore, we selected Prolific as a platform that the majority of workers are not using as a primary source of income (Abbas & Gadiraju, 2022). The results of our interviews indicate that there exist two types of crowdworkers: Either the crowdworkers liked being able to solve tasks during their everyday internet browsing and did not adjust their browsing behavior because of the *CrowdSurfer*, or they did not like the random appearance of tasks popping up on websites. The second group of crowdworkers actively searched for tasks. They identified the relevant websites by searching in the experimental descriptions where we stated that a requirement for participation is that they frequently visit some of the mentioned websites. These two groups of workers can also be linked to the work–nonwork boundary management profiles of humans (Kossek et al., 2012). There are humans who like to integrate work tasks and non-work tasks, while there are also workers who find it difficult to set appropriate boundaries to not get interrupted. Currently, the *CrowdSurfer* design mainly serves the so-called fusion lovers (Kossek et al., 2012). They liked to surf the internet and earn money during this activity. The second group of crowdworkers still liked to execute tasks at hand. Although these participants did not have the real usage scenario as it was intended for the *CrowdSurfer*, they still saw advantages in the browser extension. We assume, that this group preferred to separate work and non-work tasks and respectively set their boundaries. Similar to the first group, they also felt more in the context when providing feedback and thought that it is more seamless and less effort to use the extension to provide feedback than doing it via a survey. However, the additional search process for tasks might have confused them and increased their invisible work.

Therefore, we think it would be desirable to address both types of workers in the future. To do this, the *CrowdSurfer* could offer a list of available tasks. This also has the advantage for feedback requesters that crowdworkers could be guided to new or less frequently visited websites to provide feedback.

5.6.4. Design Recommendations for Browser Extensions to Integrate Crowdsourcing Tasks in Everyday Internet Surfing

Based on the results and the implications that we discussed in the sections above, we derived six design recommendations for the design of browser extensions to integrate tasks into crowdworkers' everyday internet surfing. These insights shall help future researchers to design similar extensions for other types of tasks. Following the structure that we used to analyze the qualitative interviews, each recommendation is assigned to one of the three concepts: *Usability*, *Work Process*, and *Working Conditions* of crowdworkers.

1. *Present users an overview of the collected data (Usability).* Showing users which data is collected about them increases the transparency of the extension which in turn positively affects users' trust. Further, users can better manage their tasks and rewards.

2. *Provide support, guidance, and background information (Usability).* Participants in our study stated that they believe that they would have been able to provide even better feedback if they had more background information on the tasks or support in solving the tasks. The extension should provide users with important information about the requirements of a task and provide them support in solving the tasks.

3. *Ensure task conduction is quick and easy to limit interruption of users in daily life (Work Process).* The integration of crowdsourcing tasks only makes sense when the tasks are simple and quick to complete. When users have to spend more time than a few minutes to solve the tasks, they might feel interrupted in their actual task and refuse to do it.

4. *Offer a way to actively search for tasks (Work Process).* We learned that there exist crowdworkers who do not like waiting for tasks to pop up and prefer a list of available tasks to complete. There should be an option for these crowdworkers to actively search for and directly access available tasks whenever they are willing to work.

5. *Support on/off functionality for the browser extension (Working Conditions).* Offering users to turn off the extension when they exclusively want to surf privately on the internet is important. Users need to keep their flexibility between work and

private time. Also, it helps to increase trust in the extension as crowdworkers can turn it off when they do not want to be traced.

6. *Make all tasks voluntary and allow the rejection of tasks (Working Conditions).* Workers liked in our study that they could freely decide which tasks they want to do and were not forced to do tasks on websites they did not feel comfortable doing. Forcing workers to do tasks might reduce their willingness to participate in the tasks at all.

5.7. Limitations & Future Work

In this section, we summarize the limitations that we acknowledge in this study and connect them to future research avenues. Further, we present our vision of how feedback requesters and crowdworkers could use the *CrowdSurfer* in practice.

First of all, we caution against overgeneralizing the findings from this *CrowdSurfer* study. Our findings are limited by the self-selecting sample of participants caused by our study design and by the websites and feedback tasks that we selected. Also, our study did not present a real feedback scenario and participants could experience the *CrowdSurfer* only for seven days and not in the long term. Due to the nature of the experiment and the innovativeness of the *CrowdSurfer* as a browser extension, this was not the case for our study. However, we believe that our results already provide good indicators for the applicability to continuously collect feedback. Overcoming this limitation requires longer user studies. While we believe that new feedback tasks need to come from real feedback requesters future work needs to bring the *CrowdSurfer* to life, connect it to real feedback requesters, and investigate its effects in the wild.

Second, as mentioned in the discussion, some participants actively searched for tasks instead of waiting for them. They were in a working mode and did not want to wait for tasks. The *CrowdSurfer* did not provide functionalities for these users to directly access the tasks at hand. Future work should derive two actions: First, (1) investigate how the feedback differs between crowdworkers who actively searched for tasks and crowdworkers who did not. Second, the *CrowdSurfer* should be designed to (2) allow users to find tasks easily and simplify the search for tasks. To do so, the *CrowdSurfer* could provide a list of all available tasks. Further, this list would also simplify the feedback collection for less frequently visited websites.

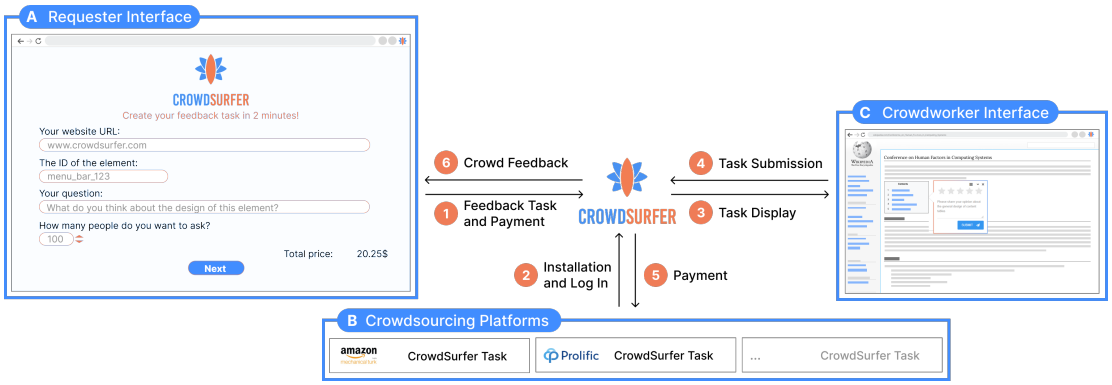


Figure 5.8.: The *CrowdSurfer* process as we envision it for implementation in practice. 1) First, requesters create tasks that are added to the *CrowdSurfer* database. 2) Crowdworkers are continuously recruited to install and set up the extension. 3) Available tasks are then displayed on the websites whenever a crowdworker accesses them. 4) The submitted task is stored in the *CrowdSurfer* database. 5) The crowdworkers are paid via weekly bonuses using the initial installation tasks. 6) The received feedback is published to the requester interface.

Third, one core element of our study was the assessment of feedback quality. We recruited crowdworkers with UI design experience to evaluate the collected feedback. Of course, this might lead to different results than asking the actual website designers and developers how valuable they perceive the feedback. Still, this method was already applied in similar studies (e.g., Haug, Benke, and Maedche, 2023) and as the feedback in both treatments was evaluated in the same way, we believe that the results are overall still valid and comparable. In further studies, the feedback could be analyzed with the help of actual feedback requesters. We assume that these feedback requesters might have different expectations regarding the feedback than the quality dimensions that we analyzed. Further, a clear explanation of feedback requesters' requirements and the benefits of real in situ user feedback compared to traditional survey-based feedback would help to design further crowd-feedback systems.

Finally, we designed the *CrowdSurfer* to conduct tasks for design feedback. However, on crowdsourcing platforms, there are multiple other types of tasks available (e.g., matching, labeling, idea creation, captioning). For being a potential tool to conduct crowdwork, the *CrowdSurfer* should allow further task types. In particular, task types that rely on internet usage and context seem prone to the *CrowdSurfer* application. Future research should investigate which task types are applicable to present with the *CrowdSurfer* and beneficial for both the task requesters and the crowdworkers.

5.7.1. CrowdSurfer Implementation Concept

While our *CrowdSurfer* field study aimed to replicate a realistic feedback scenario, there are several elements to be considered when bringing the *CrowdSurfer* into the real world beyond an experimental use case. In this section, we describe a potential implementation concept for the usage of the *CrowdSurfer* in a real-world setting for the crowdsourcing of design feedback.

Overall process

Figure 5.8 presents the structure and usage flows for feedback requesters and providers (crowdworkers). There are two main intertwined processes within the *CrowdSurfer* application, one for the feedback requesters and one for the feedback providers. First, feedback tasks need to be created by requesters, and, second, they need to be fulfilled by crowdworkers. For the first process, feedback requesters visit the *CrowdSurfer* website (A) and can, in very simple user interaction, create the task by providing the URL of the respective website, the HTML ID attribute of the element on which feedback is required, and a question that feedback providers shall answer. They can, further, indicate how much feedback they want. Feedback requesters also pay via this requester interface for their feedback requests. The payment includes the rewards for the feedback providers, the service fees for the crowdsourcing platforms, and a service fee for the *CrowdSurfer* operators. The requested task and respective payment are then stored in the *CrowdSurfer* database (1). At the same time, *CrowdSurfer* tasks are continuously presented on crowdsourcing platforms (e.g., MTurk, Prolific) (B). These tasks allow the *CrowdSurfer* operator to recruit new crowdworkers, whenever they are needed. In these *CrowdSurfer* tasks, crowdworkers are asked to install the *CrowdSurfer* extension and log in with their respective crowdsourcing platform ID (2). In return, they receive a fixed payment and the task is completed. Now, whenever a *CrowdSurfer* user visits a website on which tasks are available (3), they see the feedback pop-up (C). When they submit feedback, their answer is stored in the *CrowdSurfer* database (4). As the mentioned crowdsourcing platforms allow the payment of bonuses, the *CrowdSurfer* operator will use this functionality to pay *CrowdSurfer* users for their feedback on a weekly basis (5). Feedback requesters can then access the submitted feedback via the feedback requester interface on the *CrowdSurfer* website (6).

Task assignment

The task assignment will be based on a first-come-first-serve functionality. That means when a new task is submitted with 100 feedbacks requested, the first 100 workers who access this page will see the feedback pop-up. When workers minimize a task, the task is reserved for them as long as they stay on this website. Workers can also reserve a task by setting a reminder one time for 24 hours. When they reject a task, the task is passed on to the next *CrowdSurfer* user who visits the website.

CrowdSurfer for less popular websites

We are aware that the proposed concept requires a sufficiently large and heterogeneous *CrowdSurfer* user base to work. We also acknowledge that less popular websites will not be challenged to collect a meaningful amount of feedback via this approach. A possible solution to address this problem is to notify crowdworkers whenever a task is available on a website that is similar to the one that they are currently on. This can be done via the *CrowdSurfer* panel and for example be integrated into the task list, that we mentioned in the ideas for further improvements of the *CrowdSurfer*. To do this, feedback requesters must choose a website category to which their website belongs when creating a feedback request. Thereby, the *CrowdSurfer* favors tasks on websites that are less frequently visited and still lack a larger number of feedback submissions. This allows us to make sure that also less frequently visited websites will receive feedback submissions while at the same time we can ensure that the context is always given when providing feedback. So for example, whenever the crowdworker visits a shopping website, available tasks on other shopping websites are added to the list in the *CrowdSurfer* overview panel. When these tasks are completed by the required number of crowdworkers, the crowdworker has completed this task himself, or the last visit to a shopping website was more than an hour ago, the task is deleted from the list.

5.8. Conclusion

Real user feedback is a valuable means to evaluate and continuously improve website designs. However, it often does not lead to the desired amount of feedback. Crowdsourcing of design feedback is a scalable alternative but also comes with drawbacks. In our study, we aimed to provide a seamless approach to crowdsource in situ design feedback. Besides

developing design rationales and recommendations for the integration of crowdsourcing tasks into crowdworkers' everyday internet surfing, we wanted to understand how the in situ feedback from real users differs from traditional crowd feedback on Prolific and how crowdworkers perceive the innovative approach for conducting crowdsourcing tasks. Therefore, we developed the *CrowdSurfer*, an innovative crowd-feedback system as a Chrome browser extension, based on exploratory interviews with crowdworkers. To analyze these effects, we conducted a field study over seven days in which crowdworkers could use the *CrowdSurfer* to provide design feedback on eight popular websites. We compared the resulting design feedback and quantitative answers of a post-task questionnaire with a traditional survey-based feedback collection. Further, we analyzed the resulting feedback and conducted twelve semi-structured interviews to understand the *CrowdSurfer* experience from a crowdworker perspective. Our results show that crowdworkers enjoyed our innovative *CrowdSurfer* design, felt more in the experience, perceived the effort to be lower than in a survey, and expected their feedback to be more relevant. Nevertheless, the feedback quality was lower. Our findings demonstrate the feasibility of integrating tasks into crowdworkers' everyday internet surfing. Still, they show that offering a more effortless way to provide feedback in return for a monetary reward might also have a negative impact on feedback quality. Our results motivate further investigations for the design of similar crowdsourcing tasks. Overall, we contribute with our work to enhance the feedback collection processes while improving the working conditions for crowdworkers.

6. Study V: Preference-based Personalization of Casual Microtasking for Crowdworkers

6.1. Introduction

Crowdsourcing describes the provision of tasks that are executed by a crowd of people working on problem-solving or data collection, contributing to a common goal (Jäger et al., 2019). When the crowd receives a monetary reward for their contribution, it is called crowdworking (Durward et al., 2016). Crowdworking platforms such as Amazon Mechanical Turk (MTurk) and Prolific have become very popular and important in recent years. These crowdworking platforms are used intensively for so-called microtasks, i.e., short and simple tasks such as tagging images, verifying information, or participating in surveys (Gadiraju, Kawase, & Dietze, 2014). An innovative approach for crowdworking is embedding tasks into crowdworkers' daily lives so that they are available when convenient, so-called casual microtasking. The work by Hahn et al. (2019) shows that integrating short writing tasks into the Facebook feed allows users to leverage spare micromoments during their primary work. Another advantage is that crowdworkers are in a specific context when working on a task. This can be helpful, for example, for feedback tasks (Haug, Benke, Fischer, & Maedche, 2023). Integrating these feedback tasks into websites ensures situatedness, which can increase participation rates and engagement (Hettiachchi, Wijenayake, et al., 2020). Moreover, in a crowdworking context, this can cause less perceived effort for crowdworkers as they can skip the task of searching and entering a website and getting familiar with the content (Haug, Benke, Fischer, & Maedche, 2023).

However, similar to traditional crowdworking tasks, ensuring high-quality results is a persistent problem for casual microtasking. The outcome quality could even be worse than in traditional crowdworking (Haug, Benke, Fischer, & Maedche, 2023). Known reasons for low-quality results in crowdworking are malicious workers (Gadiraju, Kawase, Dietze, & Demartini, 2015) or a lack of intrinsic motivation (Rogstadius et al., 2021). An extensive body of research has already focused on improving the outcome quality in microtasks (Wang, Ipeirotis, & Provost, 2017). A human-centered approach for ensuring a better job performance of crowdworkers is filtering for specific characteristics of crowdworkers. Established filters are crowdworkers' approval rate, the number of completed tasks, or their

nationality (e.g., Kittur et al., 2013). Extending the traditional filtering approach, personalized task recommendations based on crowdworkers' characteristics like skills, interests, or cognitive abilities can further improve outcomes (Amer-Yahia et al., 2016; Difallah et al., 2013; Paulino, Correia, Guimarães, et al., 2022).

While casual microtasking offers the advantage of integrating small tasks into the primary activities of crowdworkers, it presents unique challenges compared to traditional crowdwork. Unlike conventional methods where tasks are the main focus, casual microtasking is more intrusive and may lead to compromised task quality, as crowdworkers are unable to fully concentrate on these secondary tasks (Haug, Benke, Fischer, & Maedche, 2023). Imagine you are working on a software development task and eventually go on YouTube to search for a tutorial on how to code some feature. There, you receive a pop-up asking you about your opinion on the recommendations of YouTube. Even though you decided to answer this request, would you be able to give it your full attention and write a long paragraph about what could be improved in the YouTube recommendation algorithm? Maybe yes, maybe no, depending on your preference for switching between tasks. This shows the necessity for an innovative approach, where microtasks are personalized to align with the individual preferences of crowdworkers. Additionally, there is a lack of research on adapting microtask design according to crowdworkers' characteristics, although the task design can have a huge impact on task outcomes (Paulino, Correia, Guimarães, et al., 2022). From a scientific perspective, research is scarce on the analysis of the fit between the environment of crowdworkers and their individual characteristics. This lack of knowledge underlines the need for further investigation to optimize the casual microtasking model and enhance its efficacy in the crowdworking landscape. Overall, we argue that there is a need to research the design of casual microtasking systems that can adapt to crowdworkers' individual preferences.

In this study, we want to address these real-world challenges that crowdworkers face in casual microtasking by suggesting a solution in the form of a personalized casual microtasking system for crowdworking that supports effective and humane microtasking. Specifically, we leverage the person-environment (P-E) fit theory as kernel theory guiding our design. According to the P-E fit theory, humans are more satisfied with a job and perform better when their abilities meet the demands of their environment or when the supplies of the environment address their needs (Cable & Edwards, 2004; Caplan, 1987). The P-E fit

theory is frequently applied in information systems (IS) research to analyze antecedents of job performance and satisfaction (Ayyagari et al., 2011; Tams et al., 2018; Venkatesh et al., 2017). In our study, we mainly focus on the fit between the crowdworkers' preferences and the job characteristics, the so-called person-job (P-J) fit as the key aspect of P-E fit. Thus, we argue that casual microtasking systems which adapt to the individual preferences of crowdworkers should impact job performance positively. Therefore, we seek to answer the following two research questions:

RQ1: *How to design a preference-based personalized casual microtasking system to increase job performance?*

RQ2: *How do preference-based personalizations in casual microtasking systems affect job performance?*

In this paper, we connect a design science study with two experimental studies. First, we follow a theory-guided design science research (DSR) approach and propose the preference-based personalized casual microtasking system *MyCrowdSurfer*. Building on existing knowledge described by the P-E fit theory and literature on preferences, such as polychronicity and altruism, we derive requirements for a personalized crowdworking system that adapts task designs to crowdworkers' preferences. We implement a running software prototype in the form of a browser extension that is able to integrate microtasks into crowdworkers' daily internet surfing. This extension can adapt the task design to crowdworkers' polychronicity and social preferences to provide a personalized task experience.

Second, we analyze the effect of preference-based personalization in casual microtasking systems on the job performance of crowdworkers. We perform two large-scale field experiments with crowdworkers actively engaging on the crowdsourcing platform Prolific. In the studies, crowdworkers need to complete an artificial primary task, that asks them to find information on Wikipedia. In parallel, they can work on a bonus task, that is providing alt-tags for images on these Wikipedia pages to increase the accessibility of Wikipedia. The participants must install the *MyCrowdSurfer* browser extension and use it for seven days to work on these microtasks. After seven days, they participate in a questionnaire that asks them about their perceptions of the person-job fit, job satisfaction, and job performance. We will also quantitatively analyze the task results in terms of alt-tag quantity, length, and relevance.

With this study, we aim to deliver knowledge on the effects of preference-based personalization in casual microtasking as well as prescriptive knowledge in the form of requirements and an instantiation of a personalized casual microtasking system that supports integrating microtasks in crowdworkers' daily lives under consideration of individual preferences. Our work contributes to the body of knowledge in the field of crowdworking by investigating the effect of personalization of casual microtasks on job performance. We show how personalization according to crowdworkers' polychronicity and altruism affects the resulting job performance in terms of quantity and quality of outcomes. We also provide evidence that under certain circumstances personalization can be counterproductive. Further, we contribute by proposing prescriptive knowledge for designing preference-based casual microtasking systems that enable the effective integration of microtasks into crowdworkers' daily lives under consideration of individual polychronicity and altruistic preferences.

The remainder of this paper is organized as follows. First, we introduce the key underlying concepts of our study, focusing on personalization and crowdworking systems to define the relevant constructs of interest in our research and provide a short overview of related work. Subsequently, we present our kernel theory and derive two requirements for the design of a personalized casual microtasking system. Based on that we present the design of *MyCrowdSurfer*. Next, we describe two studies in which we applied the instantiated design to investigate the effect of personalization on crowdworkers' job performance. Finally, we summarize our findings, discuss, the theoretical and practical contributions, and critically reflect on the limitations. Based on that, we provide ideas for future research.

6.2. Conceptual Foundations & Related Work

In this section, we explain the underlying concepts of our study and summarize related research. We start by explaining the term *personalization*. This is followed by a deep dive into personalized crowdsourcing systems, also explaining crowdworking, incentives for crowdworkers, and casual microtasking. Finally, we define the research gap that we address in this study.

6.2.1. Personalization

The Cambridge Dictionary (Cambridge University Press & Assessment, 2023) defines personalization as "the act of making something suitable for the needs of a particular person".

While a significant body of research exists on the topic of personalization in IS and related fields, there is limited consensus regarding its conceptualization. Personalization is used interchangeably with terms such as *customization*, *adaptation*, or *individualization*. For our research, we build on the generic definition of personalization as a process that alters a system's functionality, interface, information access, content, or distinctiveness to enhance its personal relevance for an individual (Fan & Poole, 2006).

The process of personalization can be distinguished into two major activities: First, information about the user (characteristics, traits, preferences, states), as well as the associated task and context, is collected, and a user model is created. Second, leveraging the user model the personalization of the system is performed. Personalization of the system can focus on different dimensions, e.g., the system's design (interaction), the system's content, or the task. Furthermore, the way personalization is triggered is critical (e.g., system/user-invoked, time, location). Based on Sundar and Marathe (2010), we only understand system-initiated processes as personalization and call user-initiated processes *customization*. Besides the anticipated positive effects of personalization on users, personalization also has drawbacks. Personalization comes with a privacy challenge as users need to disclose personal information for personalization purposes, leading to a personalization-privacy tradeoff (Awad & Krishnan, 2006).

In the field of IS, personalization plays a major role when researching online consumer behavior. Murthi and Sarkar (2003) developed a comprehensive overview of research on personalization in the management sciences, focusing on the personalization process, including the technical issues for personalizing IS, and research on the effects of personalization on the firm strategy and performance. This includes also effects on antecedents of firm performance, like the consumer's decision process. In the domain of IS, the vast majority of existing studies in the context of personalization address the second stream and investigate how personalization affects the firm performance, including constructs like customer loyalty, consumer decisions, and adoption of recommendation agents. For example, S. Y. Ho and Bodoff (2014) found that user attitudes towards web personalization agents are shaped by the number of items they sample and how deeply they think about each item. This attitude subsequently affects their behaviors in terms of further item sampling and selection. Thereby, this study offers valuable insights for online merchants on managing web personalization to optimize advertising and sales revenues. Zhang et al. (2011) found

that personalized product recommendations with higher quality significantly reduce the cost of product screening for customers and enhance the quality of their decision-making while shopping online. This can positively impact customer repurchase intentions and shows a strong connection between well-tailored personalized product recommendations and increased customer loyalty in electronic markets. There is a multitude of further studies analyzing personalization mechanisms to optimize online consumer behavior like the timing of adaptive web personalization (S. Y. Ho, Bodoff, & Tam, 2011), the effect of content relevance, self-reference, and goal specificity (Tam & Ho, 2006), or the decision of the correct personalization strategy (Thirumalai & Sinha, 2013).

While in online retail, personalization can be seen as a competitive advantage, its application in other contexts is rather scarce. The study by Eichler and Dostál (2012) is one of the exceptions. The study investigates how a personalized adaptation of the user interface based on users' activity influences user experience and productivity. Also, Leung et al. (2023) investigated personalization in a different field. They find that gamification is effective when the design provides personalized feedback matching learners' goal orientations. However, a one-size-fits-all approach to gamification can be counterproductive, suggesting that successful gamification requires careful consideration of individual learner traits.

To summarize, the existing body of research in the domain of IS personalization primarily focuses on consumer-centric applications, demonstrating its effectiveness in enhancing online retail experiences through increased engagement, more efficient decision-making, and higher consumer satisfaction (S. Y. Ho & Bodoff, 2014; S. Y. Ho, Bodoff, & Tam, 2011; Tam & Ho, 2006; Thirumalai & Sinha, 2013; Zhang et al., 2011). However, there is a notable gap in the exploration of personalization in work-related contexts. While the positive impacts observed in consumer settings are promising, they cannot be directly translated to professional environments due to the distinct boundary conditions of these settings. In the workplace, personalization extends beyond immediate decision-making or purchasing actions. It holds the potential to significantly influence long-term outcomes, such as employee satisfaction and engagement, by tailoring the work environment to individual preferences and needs. Compared to online retail, personalization in a work setting is also usually only secondary to the primary motivator of financial compensation, yet its interplay with financial incentives is complex and requires thorough investigation. Understanding this dynamic is crucial, as it could redefine employee motivation and productivity

in the modern workplace. Given these considerations, we argue for a dedicated exploration of the impacts of personalization on work outcomes, which could uncover new insights into employee engagement and efficiency. The following section summarizes the foundations of personalized crowdsourcing systems and presents existing research in this emerging field.

6.2.2. Personalized Crowdsourcing Systems and Microtasking

Crowdsourcing is a form of digital work that uses a large undefined group of people to solve tasks (Durward et al., 2016; Howe, 2008). Paid crowdsourcing, also called crowdworking (Durward et al., 2016) can be done via online platforms like Prolific and MTurk. These platforms serve as connections between job providers and crowdworkers across the world and usually offer short tasks, so-called microtasks, that only take a few minutes or even seconds to be completed (Kittur et al., 2013). Usually, these tasks don't require specific skills and are repetitive like labeling tasks, transcriptions, or surveys (Deng, Joshi, & Galliers, 2016). The tasks can be conducted directly via the platform, using additional survey platforms, or via dedicated crowdsourcing systems. Crowdsourcing systems facilitate the outsourcing of tasks to a broad online community, offering a versatile approach to task completion. Crowdsourcing systems usually have four key components: user management, task management, contribution management, and workflow management (Hetmank, 2013). Also, crowdsourcing systems can be differentiated by how they derive value from contributions and how they differentiate between contributions (Schader et al., 2012). In the following, we will include research on crowdsourcing and crowdworking. However, our study only focuses on crowdworking.

A new crowdsourcing concept is casual microtasking, which was introduced by Hahn et al. (2019). Casual microtasking is a type of crowdsourcing where microtasks are seamlessly integrated into other online activities that users are primarily engaged in. This integration allows to leverage spare micromoments and also enables to offer tasks to crowdworkers when they are already in the right context for the task. Goncalves, Hosio, et al. (2015) show the potential of context to motivate participation in ubiquitous crowdsourcing tasks. Therefore, situatedness and context have the potential to increase participation rates and engagement. Context can also lead to less perceived effort for crowdworkers and can improve certain parts of task outcomes, such as making design feedback more real (Haug, Benke, Fischer, & Maedche, 2023). Haug, Benke, Fischer, and Maedche (2023) developed

the *CrowdSurfer*, a browser extension to integrate microtasks into crowdworkers' everyday internet surfing to leverage the context for specific tasks where context could be beneficial. In this study, we build upon their system which is publicly available on GitHub.

The incorporation of casual microtasking within crowdsourcing systems suggests a shift towards more flexible and dynamic work structures. The evolving focus of crowdsourcing research towards greater flexibility and adaptability is evident through the emergence of a new research stream on personalized crowdsourcing, which has seen significant growth in recent years. Personalized crowdsourcing systems use data related to the target crowdworker to profile them and exploit a user model to filter, recommend, or adapt crowdsourcing tasks (Naudet & Lykourentzou, 2014). Naudet and Lykourentzou (2014) discuss in their paper the use of personalization in crowdsourcing and provide a foundational overview. However, in recent years, many new approaches to personalized crowdsourcing have arisen. In Table 6.1, we provide an overview of related research in the field of personalized crowdsourcing. For the personalization, the most used characteristics of crowdworkers are interests (Alsayasneh et al., 2018; Amer-Yahia et al., 2016; Difallah et al., 2013; Wang, Yang, et al., 2022), skills (Alsayasneh et al., 2018; Kurup & Sajeev, 2018; Wang, Yang, et al., 2022; Wecker, Schor, Raziel-Kretzmer, et al., 2020; Wecker, Schor, Elovits, et al., 2019), and cognitive abilities or cognitive styles (Hettiachchi, van Berkel, Kostakos, & Goncalves, 2020; Paulino, Correia, Barroso, & Paredes, 2023; Paulino, Guimaraes, et al., 2023). The user model that is used for the personalization can be generated via tests, existing user profiles, or task fingerprinting, thus analyzing crowdworkers' behavior in previous tasks (Paulino, Guimaraes, et al., 2023). For example, Alsayasneh et al. (2018) ask crowdworkers to select what combination of tasks they would prefer to work on. They use the results to assign crowdworkers to personalized task compositions. Paulino, Guimaraes, et al. (2023) use tests and task fingerprinting to assess crowdworkers' cognitive abilities, specifically executive functions like cognitive flexibility. In a case study, they demonstrate that these methods, combined with a deep learning model, can effectively predict task performance with 95% accuracy.

Table 6.1.: Overview of personalized crowdsourcing research

Study	Type of Study	Assessment	Personalization Characteristic	Personalization Dimension	Results
Alsayasneh et al. (2018)	Algorithm and experimental user study	User selection	Interests (type of task) and skills	Task assignment	Personalization enhances worker experience
Difallah et al. (2013)	Artifact (crowdsourcing system) and experimental user study	Social network profile	Interests (e.g., likes on social media)	Task assignment	Personalization leads to higher task accuracy
Hettiachchi, van Berkel, Hosio, et al. (2019)	Experimental user study	Cognitive test performance	Cognitive abilities	Task assignment and recommendation	Personalization improves task accuracy
Hettiachchi, van Berkel, Kostakos, and Goncalves (2020)	Artifact (crowdsourcing system) and experimental user study	Cognitive test performance	Cognitive abilities	Task assignment and recommendation	Personalization improves task performance
Kurup and Sajeev (2018)	Algorithm and experimental study (evaluation of algorithm performance)	Skill taxonomy mapping	Skills	Task recommendation	Expected: personalization allows new workers to find matching tasks faster
Organisciak et al. (2015)	Experimental user study	Profiling tasks	Taste	Task recommendation	Personalization improves subjective task outcomes
Paulino, Correia, Guimarães, et al. (2022)	Case study	Cognitive test performance	Cognitive styles	Task design	Personalization improves task matching for better task outcomes
Paulino, Guimaraes, et al. (2023)	Case study	Cognitive test and task fingerprinting	Cognitive abilities	Task design	Personalization improves task accuracy effectively
Wang, Yang, et al. (2022)	Algorithm and experimental study (evaluation of algorithm performance)	Historical repositories of platform	Interest, skills (preferences and technical abilities)	Task recommendation	Personalization leads to less effort for crowdworkers and more efficient matching
This study	Design study and field experiment	Questionnaire	Polychronicity and social preferences	Task design	Expected: personalization improves job performance

Regarding the personalization dimension, recent research has mainly shown that personalized task composition and recommending tasks according to crowdworkers' skills, interests, and abilities can improve crowdworkers' experience, task throughput, and task results (Amer-Yahia et al., 2016; Difallah et al., 2013). Geiger and Schader (2014) provide an overview of personalized task recommendations in crowdsourcing. They provide a conceptual foundation for designing personalized task recommendation mechanisms. Especially the personalization according to cognitive abilities and preferences has received much attention in recent years. Hettiachchi, van Berkel, Kostakos, and Goncalves (2020) developed a system that recommends and assigns tasks according to crowdworkers' results in fast cognitive tasks. Thereby, they can increase the task performance.

Besides personalizing the task selection phase, Wecker, Schor, Elovits, et al. (2019) and Wecker, Schor, Raziel-Kretzmer, et al. (2020) propose further ideas for personalization in different phases of crowdsourcing according to crowdworkers' characteristics like motivational messages, tutorial material, and feedback on crowdworkers' progress. Paulino, Correia, Guimarães, et al. (2022) state that "task design is one of the core aspects of the crowdsourcing process and its optimization is a priority for many requesters that want to have their tasks solved in short times and with high levels of accuracy" (p. 484). Therefore, they explore the adaptation of task designs according to information processing preferences. Their results show that UI adaptations can improve outcomes and acceptance rates of crowdworkers.

6.2.3. Research Gap

In this paper, we address a significant gap in the field of personalized crowdworking, focusing on the promising approach of casual microtasking. While casual microtasking offers unique benefits compared to traditional crowdsourcing tasks, it also presents new challenges. Unlike traditional crowdsourcing where tasks are the primary focus of the crowdworker, in casual microtasking, the task is often only the secondary task, which can potentially lead to reduced focus and quality of results.

At the same time, personalization of task recommendation, assignment, and design in crowdwork is evolving. While it is able to improve task outcomes, there is a lack of a theoretical approach to personalized crowdsourcing. Further, most personalization approaches focus on skills and interests as workers' characteristics and adapt task recommendations

and assignments, but not the task design itself. We want to tackle this research gap by designing preference-based personalized casual microtasking systems and understanding the effects of preference-based personalization on job performance following the P-E fit theory.

We argue that casual microtasking must respect workers' polychronicity to be adopted by crowdworkers. Further, previous research has shown that people exhibit social preferences and are thus also motivated by non-monetary factors like having an impact and contributing to something good. We argue that personalizing the task design of casual microtasks according to crowdworkers' individual preferences can also improve job performance.

6.3. MyCrowdSurfer - A Preference-based Personalized Casual Microtasking System

In this section, we first describe our design method and the underlying theory and requirements as well as the context of our study. We then present our artifact and the design instantiations for our requirements.

6.3.1. Design Method

For the first part of the paper, we follow the design science research (DSR) paradigm to answer RQ1 and propose a theory-driven design for personalized casual microtasking systems. In particular, we draw on the P-E fit theory as the kernel theory for our design (Caplan, 1987). We focus on P-J fit as one aspect of the P-E fit theory which posits that the fit between a person's abilities, needs, preferences, and values and a job's supplies, demands, and values affects this person's job satisfaction and job performance. We exploit this theory by deriving two requirements, instantiating them in two design instantiations that are both adaptive to two contrary preferences. We then rigorously evaluate the instantiations of our requirements in two experimental field studies (Venable et al., 2016).

6.3.2. Kernel Theory

6.3.2.1. Person-Environment Fit Theory

The person-environment (P-E) fit theory explores the interplay between individuals and their work environments (Edwards, Caplan, & Van Harrison, 1998). It posits that a fit

between the characteristics of a person and the characteristics of the work environment influences job satisfaction and job performance. Therefore, individuals seek P-E fit, broadly defined as the "congruence, match, similarity, or correspondence between the person and the environment" (Edwards & Shipp, 2007, p. 212). Previous research presented two different types of P-E fit: While supplementary fit describes the similarity of characteristics between the human and the work environment, complementary fit describes how the characteristics of the human complement the characteristics of the environment. Supplementary fit is achieved when humans perceive a value congruence with the environment. Complementary fit can further be operationalized as need-supply fit or demand-ability fit (Kristof, 1996). Both are often used to measure employees' perceived fit with their jobs, rather than their workgroups or organizations (Guan et al., 2011; Piasentin & Chapman, 2007). There are three levels of P-E fit, that individuals might search for in their workplace: person-organization fit, person-job fit and person-group fit (Kristof, 1996). Person-organization fit (P-O fit) deals with the fit between the person's values, beliefs, and goals with the organization's culture. Person-group fit (P-G fit) deals with the match between the person and the workgroup. A high fit leads to fewer conflicts and better collaboration (Kristof, 1996). Person-job (P-J) fit deals with the fit between the person's abilities, preferences, skills, and needs with the requirements and offerings of the job (Sekiguchi, 2004). In our study, we focus on the P-J fit as this fit is the most important aspect of crowdwork.

P-J fit is a well-researched concept, especially in the context of recruiting and job engagement (Chen et al., 2014; Sekiguchi, 2004; Warr & Inceoglu, 2012). Existing studies that combine crowdsourcing and P-J fit only focus on skill matching or simple selection mechanisms according to rating reputation or approval rate (Buettner, 2015). Therefore, Buettner (2015) calls for more empirical and design-oriented research on P-J fit mechanisms in crowdsourcing.

6.3.2.2. Polychronicity

Polychronicity is considered the preference for handling multiple tasks at once, also called multitasking (König & Waller, 2010). It encompasses their inclination to engage in concurrent activities, such as performing two or more tasks simultaneously or switching attention among multiple tasks. The term *polychronicity* is used to describe peoples' preferences for multitasking, while the actual behaviors, rather than attitudes, should be

termed *multitasking* (König & Waller, 2010). Those with higher polychronic tendencies, often referred to as "polychrons", exhibit a preference for multitasking and are more comfortable with interruptions and switching activities. Those with lower polychronic tendencies, known as "monochrons", lean towards monotasking, where tasks are executed sequentially. Monochrons are known for strict planning, concentrating on and prioritizing tasks (Kaufman-Scarborough & Lindquist, 1999). There are two types of multitasking: Dual-tasking refers to performing two activities simultaneously, such as driving a car and listening to music (Huxhold et al., 2006). Task-switching means that the attention is allocated among multiple tasks before completing any task compared to completing the tasks sequentially (Koch, Gade, et al., 2010; Koch, Poljac, et al., 2018; Monsell, 2003). Polychrons are humans who have a preference for task-switching, dual-tasking, or both. However, most research on polychronicity considers polychronicity mainly as the preference for task-switching, which can also be seen by the focus on task-switching in existing scales to measure polychronicity (Bluedorn et al., 1999; Kaufman et al., 1991; Lindquist & Kaufman-Scarborough, 2007). In this study, we also only consider task-switching when talking about multitasking. Polychronicity is considered to be a relatively stable individual difference (Howard & Cogswell, 2023).

Lascău, Gould, Cox, et al. (2019) investigate the multitasking behavior of crowdworkers. In their study, they provide recommendations for crowdworking platform owners and task designers on how to design for crowdworkers' preferences. Their recommendations for task designers are rather broad (e.g., "Pay well") and mainly have the goal of not forcing crowdworkers into a multitasking behavior when they prefer monotasking. We argue that there is a need to better understand how task designs can be adaptive to crowdworkers' preferences like polychronicity.

There is also empirical support that investigated the P-E fit perspective of polychronicity. Hecht and Allen (2005) and Kirchberg et al. (2015) identify a connection between polychronicity values and workers' well-being and job satisfaction following the P-E fit theory. Asghar, Tayyab, et al. (2021) and Asghar, Gull, et al. (2020) utilize the P-E fit theory to research the effect of polychronicity on turnover intentions, and job performance. The studies were conducted in the context of service jobs, where polychronic workers might experience a greater fit due to the required multitasking. König and Waller (2010) suggests that the effect of polychronicity on job performance depends on the fit between the

demands of the task and the abilities of the worker, calling for more empirical research on P-E fit and polychronicity. In our study, we aim to investigate polychronicity and the P-J fit in the context of crowdworking, especially casual microtasking, which usually requires multitasking. To optimize job performance, we build a personalized casual microtasking system that adapts the task design to crowdworkers' polychronicity.

Polychronicity is a preference that is nowadays beneficial in many jobs. Crowdworkers tend to multitask although research showed that they do not all prefer multitasking (Lascău, Gould, Cox, et al., 2019). However, multitasking also has its downsides. When switching from a primary task to a secondary task and back, workers always need some time to immerse in the task again. This time depends on multiple factors (McFarlane & Latorella, 2002). Also, when monotaskers are forced into a multitasker setting and the other way around, work performance will decrease.

6.3.2.3. Social Preferences

An important factor that can impact the performance of crowdworkers is social preferences. While the term preference, especially in the crowdsourcing context, is often used to describe crowdworkers' interests or types of tasks they like doing (Amer-Yahia et al., 2016), social preferences in economics relate to how people behave with others. Research in this field has shown that people are not entirely self-interested, as the concept of *homo economicus* (Levitt & List, 2008) assumes. Instead, people exhibit social preferences when they care not only about themselves but also about the well-being and profit of others (Charness & Rabin, 2002; Fehr & Fischbacher, 2002; Levitt & List, 2007). Experiments revealed that these preferences occur in several ways. They can be observed in people's intention to increase social surplus but also appear when people have a desire to reduce the differences between their own payoffs and those of others (Charness & Rabin, 2002). The most frequent and important social preferences include altruism, inequity aversion as a fairness preference, positive and negative reciprocity, as well as trust (Fehr & Fischbacher, 2002; Fehr & Schmidt, 2001; Levitt & List, 2007). The concept of the *homo economicus* who is not only interested in its own benefits like money, matches well with research on crowdwokers' motivation. Multiple studies investigated why crowdworkers work, how they can be incentivized, and how intrinsic and extrinsic motivation are related (Deng & Joshi, 2016; Law et al., 2016; Rogstadius et al., 2021). Buettner (2015) listed altruism, a social

preference, as one of the most frequently mentioned motives of individuals participating in crowdsourcing activities. Social preferences impact individuals' motivation to exceed effort and can partly compensate for monetary incentives (DellaVigna & Pope, 2018). Furthermore, field experiments revealed that employees' social preferences towards their employer also matter at work and impact their working behavior (DellaVigna, List, et al., 2022). Therefore, we argue that personalization based on social preferences as a way to increase crowdworkers' intrinsic motivation is a promising approach to improving crowdworkers' job performance.

6.3.3. Requirements

P-E fit theory suggests that humans try to achieve a complementary or supplementary fit between their characteristics and the work environment (Cable & Edwards, 2004; Guan et al., 2011). This fit can address a multitude of characteristics, such as skills, interests, values, preferences, needs, and work style. Previous research on personalized crowdsourcing focused mainly on apparent crowdworker characteristics like skills and interests that can easily be matched with task requirements. In this study, we focus on the more subtle characteristics of crowdworkers, namely their preferences. While it is difficult for crowdworkers to identify tasks that match their more subtle preferences like work style, working on tasks that do not fit crowdworkers' preferences might reduce their satisfaction and job performance. Further, we argue that although personalized recommendations and assignments of tasks to workers have been proven to be beneficial, it is worth investigating personalized task designs. By making small adaptations to the task design, the same task might fit two contrary preferences and thereby lead to higher satisfaction and better task outcomes for all workers. Paulino, Correia, Guimarães, et al. (2022) and Paulino, Guimaraes, et al. (2023) showed that adapting the task design to crowdworkers' characteristics can lead to better task outcomes and higher acceptance rates. Therefore, we will propose two requirements for preference-based personalized casual microtasking leveraging the person-environment fit theory. We identified two types of preferences that are crucial for casual microtasking and therefore, need adaptation mechanisms.

Casual microtasking is a very intrusive form of crowdsourcing as it integrates crowdsourcing tasks into crowdworkers' everyday internet surfing. These crowdsourcing tasks therefore have the potential to interrupt or distract crowdworkers during other primary

activities. While some crowdworkers enjoy this form of work, others feel interrupted and would prefer to work only when it is convenient to them (Haug, Benke, Fischer, & Maedche, 2023). Preferring to work on multiple things at the same time, also called multitasking, is characteristic of polychrons. Monochrons, on the other side, prefer to finish one task before starting with the next one, also called monotasking. Multitasking behavior and polychronicity are known to have direct impacts on factors like well-being and job performance. When multitasking, polychrons achieve higher values for well-being than monochrons (Kirchberg et al., 2015). Other studies applying P-E fit theory have already proposed that polychronicity directly affects job performance (Asghar, Gull, et al., 2020). To achieve a P-J fit for both, polychrons and monochrons, we need contrary task designs. As casual microtasking is a task type that usually promotes multitasking behavior, we see great potential when adapting the task design to the polychronicity of the current user. Consequently, we argue that the crowdworkers perform better when the casual microtasking system fosters their preferred working style. Therefore, we articulate the following first requirement for the design of casual microtasking systems:

Requirement 1 (REQ1): Casual microtasking systems should be personalized to crowdworkers' polychronicity in order to increase the fit between crowdworkers' multitasking preferences and the microtasks' required behavior leading to higher job performance.

There exists much research on the relationship between incentives, motivation, and personal characteristics of crowdworkers. Research on incentives for crowdworkers often mentions aspects like "making an impact" or "doing something good" as motivational factors (Deng, Joshi, & Galliers, 2016). It is clear that crowdworkers in general are not motivated by monetary incentives alone. Like any other human, crowdworkers do not only care for their own benefit but also for the well-being of others (Charness & Rabin, 2002; Fehr & Fischbacher, 2002; Levitt & List, 2007). For example, Deng and Joshi (2016) show that task significance and meaningfulness are important motivators for crowdworkers. This shows that crowdworkers care about the impact of their work. However, all humans, and also all crowdworkers, differ in their social preferences. As described in the P-E fit theory, to perceive a fit between an environment and yourself, the characteristics of the environment must be visible. Consequently, we argue that without changing the microtask itself or its purpose, the same task can highlight different goals that can address different social preferences. For example, a simple labeling task could highlight how these labels help to make

the resulting algorithm fairer, but it could also focus on the benefits of the crowdworkers themselves. Consequently, we assume that adapting the highlighted social preferences to the crowdworkers' social preferences can increase the perceived fit between the task and the crowdworkers' preferences. Therefore, we articulate our second requirement:

Requirement 2 (REQ2): Casual microtasking systems should be adaptive to crowdworkers' social preferences in order to increase the fit between crowdworkers' preferences and the microtasks' supplies leading to higher job performance.

6.3.4. Design Instantiation

To evaluate our design we used the existing casual microtasking system *CrowdSurfer* to integrate crowdsourcing tasks into crowdworkers' daily internet surfing. We adapted the existing system design to instantiate the two requirements. Further, we needed an artificial microtask that we could use as context for the evaluation. For personalization according to social preferences, the context is not trivial, as it can trigger social preferences, such as a prosocial goal. In the following we will first explain the context for our casual microtask, then we will give an overview of the general design of the casual microtasking system, and finally, we will explain, how the two requirements were instantiated in the design.

6.3.4.1. Context

As casual microtasking is especially beneficial for tasks that profit from the user being in a specific context, we were looking for a task that requires context. At the same time, we needed a task that has the potential to trigger any social preferences. We decided on the task of providing alt-tags for images on Wikipedia. Alt-tags (also called alternative texts) are descriptions of images that make them accessible (e.g., for people with visual impairments) as they can be read by screen readers. The lack of high-quality alt-tags is a persistent problem in web accessibility and can not fully be solved by automatized solutions relying on artificial intelligence (Stangl, Morris, & Gurari, 2020; Stangl, Verma, et al., 2021). For writing good alt-tags, the context of the image must be considered (Kreiss et al., 2022). Therefore, this is a great task for demonstrating the benefits of casual microtasking. At the same time, the purpose of the task has the potential to address social preferences, especially altruism, as it has a prosocial impact on others.

6.3.4.2. The MyCrowdSurfer System

For this study, we build upon the existing casual microtasking system (Haug, Benke, Fischer, & Maedche, 2023). This system is developed as a Google Chrome extension that can be used to integrate microtasks into crowdworkers' everyday internet surfing and, thereby, leverage spare micromoments or take advantage of crowdworkers being in a specific context when doing tasks. The *CrowdSurfer* design was developed and evaluated in a previous study on the collection of crowd feedback on website designs (Haug, Benke, Fischer, & Maedche, 2023). The *CrowdSurfer* recruits and pays participants via existing crowdworking platforms like Prolific or MTurk. The artifact itself consists of two main elements: The panel to manage the task (Figure 6.1) and pop-ups that appear whenever a task is available on a website (Figure 6.2). The panel that is shown in Figure 6.1 opens when clicking on the icon in the extensions bar (1). It can, for example, show how many tasks were already conducted by the crowdworkers (4) and give general instructions. The panel also provides a link to redo the tutorial (3) and a button to turn the extension off whenever crowdworkers want to switch to a private mode (2). This is much appreciated by crowdworkers and helps them to overcome privacy concerns (Haug, Benke, Fischer, & Maedche, 2023). The further functionalities are specific to our study and will be explained in the next section. The other part of the *CrowdSurfer*, the pop-ups, are the tools to submit answers to microtasks (Figure 6.2). They are usually attached to an element on a website, contain a specific task instruction or question, and can be minimized or rejected. When minimized, small icons still show the availability of tasks. The casual microtasks can last for a fixed period or be unlimited. To learn how this casual microtasking system can be applied in practice and to understand the related payment and task assignment processes, we refer to the previous study by Haug, Benke, Fischer, and Maedche (2023).

We used the general setup of the casual microtasking system and adapted it to our specific context of collecting alt-tags for images on Wikipedia. We call our personalized version of the system *MyCrowdSurfer*. In practice, pop-ups would show up for all images that require an alt-tag on Wikipedia. In a real-world scenario, where the extension is available for the long term and offers various tasks, crowdworkers would accidentally find the tasks when searching for something on Wikipedia. When they enter a Wikipedia page, they usually have a primary task in mind, e.g., finding specific information. When they see a task pop-up they get interrupted and are tempted to multitask. For monochrons, it would

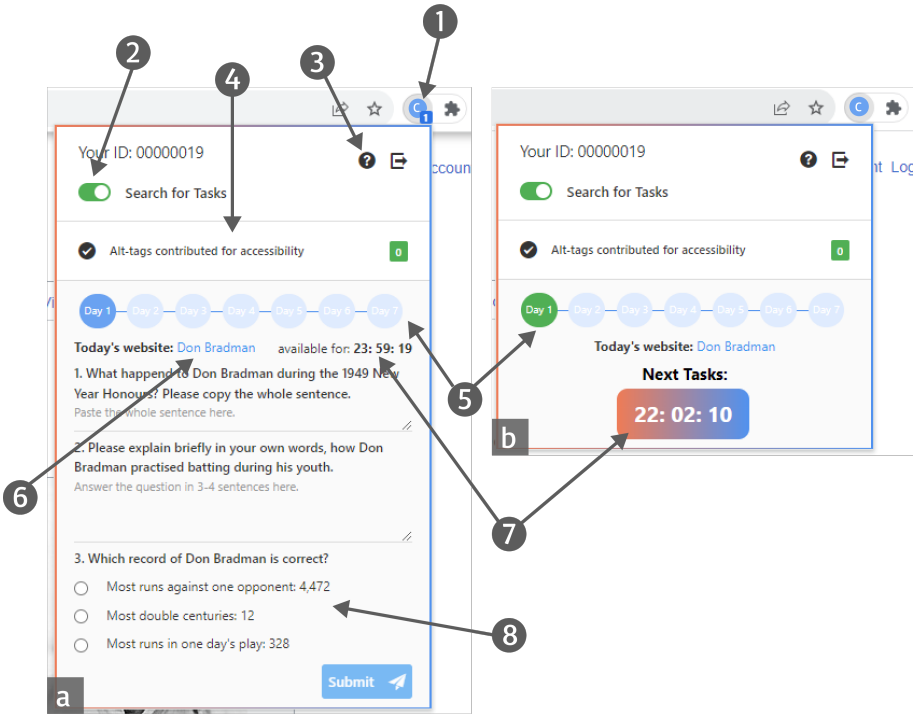


Figure 6.1.: Screenshot of the panel of our casual microtasking system

probably be more convenient when they can first finish their primary task and can then work on the casual microtask and provide alt-tags for images.

Design adaptation for experimental study

Conducting an experimental study brings some limitations that made it necessary to adapt the intended design of the casual microtasking system for alt-tags. First, we need to limit the duration of the study and make this transparent for participants. While in a real-life setting, the collection of alt-tags could be unlimited as Wikipedia is continuously changing and new images are added every day, we needed a predefined end in our study to analyze the results. Second, in an experimental study, as we planned to do, we saw the risk that crowdworkers would actively search for tasks to earn more money. In such a case, we would not be able to track if they were monotasking or multitasking. Third, we needed to control the task as much as possible to make the results comparable between participants. To be able to control the multitasking behavior of crowdworkers and counteract the active search for tasks, we had to create an artificial main task. The goal of the main task was to direct participants to websites that offer alt-tag tasks. Providing alt-tags is then only presented as a voluntary bonus task. This has the benefit that we can limit the number of available alt-tag tasks to the web pages that are included in our main task. Further, as we

are in control of the primary task, we can also track their interaction with the primary task to get a better understanding of their task-switching behavior. The main task provided a link to a specific topic (6) and asked crowdworkers three questions about this topic that could all be answered by using the Wikipedia page (8). The goal of this task is to mimic a real-life scenario, where users visit Wikipedia to search for some specific information and then see the alt-tag tasks. We randomized the order of the tasks for each crowdworker and decided that participants had 24 hours to complete each task. As we decided to limit the study to seven days, we included seven different main tasks in the system. In the panel, we show an overview of the main tasks in the form of a timeline (5). This timeline shall provide transparency to crowdworkers about how many main tasks they have solved and how many are left. The panel also includes a timer that always shows how much time is left until the current task is finished and a new main task will appear (7). Finally, to make results more comparable between crowdworkers and to get enough data points, we showed tasks for all images on the task webpages, whether or not these images already have an alt-tag.

6.3.4.3. Instantiation regarding Polychronicity

Casual microtasking in general enforces multitasking behavior which might not be appreciated by all crowdworkers as studies show that there is also a vast amount of crowdworkers who tend to prefer monotasking (Lascău, Gould, Cox, et al., 2019). As explained in Section 6.3.2.2, we focus in this study only on the task-switching aspect of multitasking. In our context, multitasking means that crowdworkers interrupt their primary task (searching for specific information on Wikipedia) to work on the secondary task (providing alt-tags for images). Monotasking behavior means that crowdworkers first finish the primary task before working on or even thinking about the secondary task. This means the secondary task should in the best case only appear after the crowdworker is finished with the primary task. While in a real-world scenario, it would be a task for itself to automatically detect when someone has found the desired information on Wikipedia and has the capacity for a new task, our setup with the artificial main tasks facilitates this.

For our task design, this means that in the multitasking instantiation, participants could provide alt-tags before, during, or after answering the questions of the main task about the Wikipedia page. However, in the monotasking design, participants could only work on

the alt-tags after they submitted their answers for the main task, thus they were not able to multitask.

6.3.4.4. Instantiation regarding Social Preferences

As with any other human being, crowdworkers do not only differ in their polychronicity but also in other work-related preferences, such as social preferences, which can impact motivation at work (see Section 6.3.2.3). In our study, we focus on altruism as a social preference that describes people's intention to consider the interests of others without having only selfish ulterior motives (Andreoni et al., 2010). Thus, while some humans are rather selfish and do not care much about other human beings, others are more altruistic and care more about society and the impact of their work on others. Research revealed that altruistic incentives (e.g., in the form of donations) can affect human behavior. We chose this social preference as the creation of alt-tags serves an altruistic goal, thus making Wikipedia accessible for visually impaired users. Research showed that workers' performance in voluntary tasks can increase if they are not paid, but their earnings are donated to charity (Charness, Cobo-Reyes, & Sánchez, 2016). Other studies support this positive effect of donations as altruistic incentives on effort (DellaVigna & Pope, 2018; Imas, 2014). One way to consider altruistic preferences is to add a prosocial mission to jobs. Cassar (2018) showed that human effort increases when adding a prosocial mission to a job (donation to charity) compared to a job without a mission. Of interest for our study is that these incentives do not motivate everyone in the same way and usually depend on workers' productivity (Tonin & Vlassopoulos, 2015) or prosociality (preference for donating money and volunteering) (Cassar & Meier, 2021). Therefore, preference-based personalization of microtasking by changing the framing and presentation of a task could be of high relevance for motivating crowdworkers. In our context, we added a prosocial job mission with our alt-tag bonus task. The overall purpose of this task, collecting alt-tags to make Wikipedia more accessible, can be seen as an altruistic goal. By changing the framing and presentation we vary whether the prosocial mission of the crowdworkers' job is highlighted or not. Altruistic preference would mean that crowdworkers care much about the social purpose of the task and exert more effort to provide alt-tags when the job mission is highlighted. Selfish preference would mean that crowdworkers care much about their own monetary payoff and exert more effort when financial incentives are highlighted.

For the personalization according to the worker's altruistic preference, we were looking for a design that does not differ in the task itself, the payment structure, and the features of the casual microtasking system. These aspects could all have an impact on job performance without relating to the P-J fit. Also, we wanted to personalize the task in a generalizable way to make our results transferable to other microtasks as well. Therefore, to address REQ2, we were looking for a way to differentiate the focus of the task, without changing the task itself. We first analyzed the different components of casual microtasks to decide where casual microtasks can be personalized according to workers' preferences. We separate the components into (1) the task presentation on crowdworking platforms, (2) the task instructions and the setup, and (3) task management and feedback. In the following, we explain how we used these components to personalize our task design.

Task presentation on crowdworking platform.

Most paid microtasks start on a crowdworking platform like Prolific or Amazon Mechanical Turk. There, the task presentation usually consists of a title, a short description of the task, and payment information. Grady and Lease (2010) showed that these task characteristics already impact the task acceptance behavior of crowdworkers. Thus, in a real-world scenario, this could be the first way to personalize tasks. However, for our study, we needed a similar number of participants in all treatments and did not want to induce a bias before even starting the actual study. Therefore, we kept the task presentation on the crowdworking platform the same for both designs.

Task instructions and setup.

After agreeing to work on a task, participants are often redirected to another platform like a separate crowdsourcing system or a survey platform. There, they receive instructions for a task, sometimes they need to complete comprehension checks to show that they understood the instructions and then they can start the task. We decided to vary the instructions according to the goal we pursued in the task. In the altruistic one, the introduction emphasizes that the crowdworker is contributing to a more inclusive world and making Wikipedia more accessible so that everyone can experience it. In the selfish system, the focus is rather on the users themselves that they can earn money by doing an exceptional job and that their skills are needed.

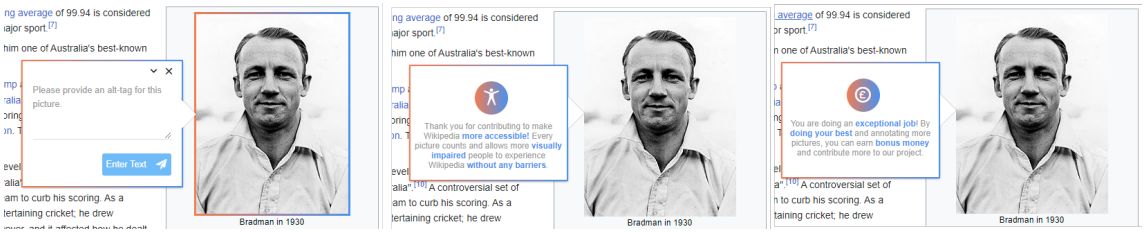


Figure 6.2.: Screenshot of the pop-ups of our casual microtasking system. Left: Task instruction. Middle: Task feedback, after submitting alt-tag to present altruism. Right: Task feedback, after submitting alt-tag to present selfishness.

Task management and feedback.

While doing the task, participants can get feedback on their performance. These feedback messages are also an element where personalization could be implemented. The effect of feedback messages has already been investigated extensively (e.g., Huang et al., 2018; Lim et al., 2021; Staiger et al., 2022). We provide feedback in two ways. First, in the panel of the extension, the number of alt-tags (altruistic) (see Figure 6.1 (4)) or the bonus money earned (selfish) is displayed to give feedback on the amount of bonus work participants already did. Second, after submitting an alt-tag the systems display different messages. While in the altruistic system the message thanks the user for helping to make Wikipedia more accessible, the selfish system tells the user that she is doing an exceptional job and is earning extra money. We ensure that both the introduction and the feedback messages have a similar length in both systems. The messages are displayed in Figure 6.2. In Table D.4 in the Appendix, we provide an overview of the exact differences between the treatments.

6.4. Experimental Studies

6.4.1. Hypotheses

We articulate one main hypothesis to guide the subsequent evaluation episodes (Venable et al., 2016).[1] Thereby, we will assess whether the instantiation of our proposed design for preference-adaptive casual microtasking systems fulfills the purpose of our design, i.e., to increase the job performance of crowdworkers who interact with a system that fits their individual preferences, like polychronicity and altruism. As outlined above, we argue that the positive effect on job performance results from the microtask satisfying crowdworkers'

[1] We preregistered the study, including the study design and the hypotheses at aspredicted.org.

needs and requiring crowdworkers' abilities. According to P-J fit, jobs that fit workers' needs, abilities, and preferences will lead to more satisfied workers and, in turn, to better performance in the job. Also, other studies found positive effects of systems that account for a higher P-J fit (Edwards, Caplan, & Van Harrison, 1998). Chilton et al. (2005) showed that a better fit between software developers' cognitive styles and the demands of the job leads to less strain and better job outcomes. Thus, there is empirical evidence of the positive effects of systems accounting for person-job fit on the performance of workers. In the context of accessibility feedback, we are specifically interested in positively impacting crowdworkers' behavior to contribute a large number and a higher quality of alt-tags. Drawing on existing research we argue that preference-adaptive casual microtasking systems will positively affect the job outcomes of the microtask.

Hypothesis 1a (1b): Crowdworkers who interact with a casual microtasking system that fits their individual polychronicity (altruism) achieve higher task performance than crowdworkers who interact with a crowdsourcing system that does not fit their polychronicity (altruism).

Besides the main hypothesis, we are also interested in understanding how job performance in casual microtasking relates to other constructs. Therefore, we will measure and report more constructs than only the job performance and aim to understand how these constructs are related.

6.4.2. Procedure

We conducted two separate studies via the platform Prolific, one for each instantiation. In study 1, we investigate the polychronicity-personalized design, while in study 2, we analyze the altruism-personalized design. Besides the different *MyCrowdSurfer* designs, both studies follow the same procedure as depicted in Figure 6.4. We will explain the procedure in the following.

Our two studies were conducted as longitudinal field studies with three parts in total: pre-screening, main task, and post-task questionnaire. In all three parts, we were using LimeSurvey for the instructions and the survey. The first part was a pre-screening survey in which participants answered questions about demographics, controls, social preferences, and polychronicity. We then analyzed the results and excluded participants who did not fit our predefined requirements regarding primary browser, English Level, and nationality.

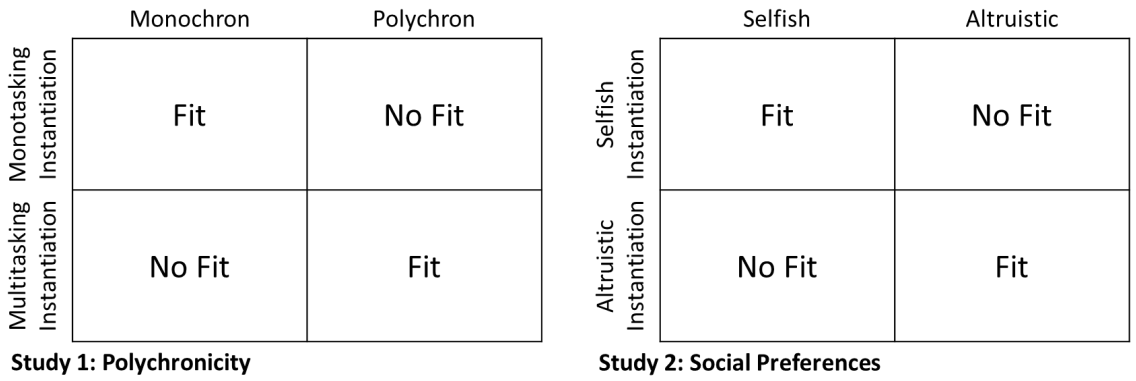


Figure 6.3.: 2x2 matrices for study 1 (left) and study 2 (right)

We also analyzed the results for their polychronicity (study 1) or altruism (study 2) as we only included workers with more extreme preferences in our study. For participants with no clear preference, we do not expect a significant effect in a personalized treatment. Therefore, we excluded participants who scored between the 0.4 and 0.6 quantil of the standardized polychronicity (study 1) or altruism (study 2) values. The remaining participants were then randomly assigned to the multitasking or monotasking design (study 1) or the altruistic or selfish design (study 2). By doing so, we received a 2x2 matrix for each study, as we could distinguish our participants by low or high values for polychronicity (study 1) or altruism (study 2) and a fit or no fit between their preference and the *MyCrowdSurfer* design (see Figure 6.3). For study 1 we used the altruistic messages in both treatments, as they represent the baseline design. For study 2, we used the multitasking setup as this is how casual microtasks are usually integrated into the daily internet surfing of crowdworkers (see Figure D.1 in the Appendix). The second part is the main part of our study, in which participants had to use the extension over seven consecutive days. The experimental task is explained in more detail in the following section. After finishing seven days, the participants were invited to complete our post-task questionnaire. In this questionnaire, they were asked about their perceptions of the *MyCrowdSurfer* design in general, and the main task and the alt-tag task separately.

6.4.3. Experimental Task

The experimental task description instructed participants to use the Chrome extension for seven consecutive days. On each day, participants received a new main task via the extension that led them to a specific Wikipedia page. On these websites, the alt-tag bonus

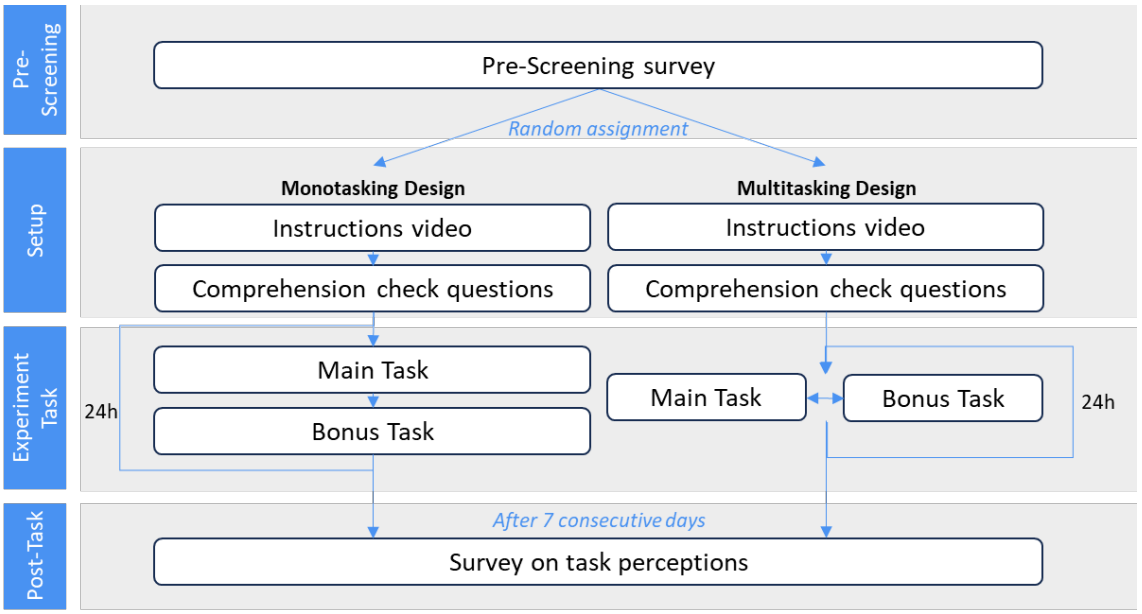


Figure 6.4.: Study procedure for the monotasking design (left) and the multitasking instantiation (right). Also, the altruistic and selfish designs both used the multitasking setup.

tasks were then either available immediately (multitasking design) or only after submitting the main tasks (monotasking design). The minimum requirements to successfully complete the second part of the study were to install the extension, complete the setup including watching a video with the instructions for the tasks, submit six of the seven main tasks, and have at least 50% of answers correct. For the seven tasks, we chose topics on which we expected the participants to have similar prior knowledge and that are not related to one of the nationalities of our participants (US, UK, South Africa). We also avoided topics that some participants might be emotional about, like politics, sports clubs, or celebrities. Finally, we aimed to cover different types of images like photos, graphs, charts, illustrations, and pages with a varying number of images. Therefore, the topics for our tasks were *Adidas*, *United Nations*, *Palomino*, *Don Bradman*, *Brazilian Carnival*, *Tyrol* and *Marketing Strategy*. Each participant received the seven tasks in a random order. Each task consisted of three questions. For the first question, participants must search the answer on the Wikipedia page and copy the respective sentence. Thereby, we could ensure that participants used the English Wikipedia page to answer the questions and see the available alt-tag tasks instead of using other websites and missing the alt-tag tasks. The second question asked them to summarize a paragraph or compare two aspects that were presented in the article in three to four sentences. Therefore we wanted the participants

to actually engage with the content of the Wikipedia article to get an understanding of the context of the images. The last question was a multiple-choice question with three answer options. With all three types of tasks, we wanted to mimic a real interaction with Wikipedia as it could happen when participants search for something on Wikipedia in real life.

6.4.4. Measures

6.4.4.1. Controls, Attention, and Comprehension Checks

In the pre-screening, we collected demographic variables such as gender, age, and education. We also asked the participants for their English level according to the Common European Framework of Reference for Languages (CEFR), for their nationality, and for their primary browser to exclude participants who have a lower English level than B2, are not from the US, UK or South Africa or do not use Google Chrome as their primary browser. In the first part, we included one attention check, in the main task, we included one comprehension check in which participants had to answer six questions about their task, and in the third part, we included three to four comprehension checks and two to three attention checks. These checks made sure that our participants were attentive while participating in our study. In the post-task questionnaire, we also asked the participants for their perceived fairness of payment based on the items of Alpar and Osterbrink (2018) and included the Situational Intrinsic Motivation Scale (SIMS) to better understand their motivation Guay et al. (2000).

6.4.4.2. Preferences

For measuring participants' polychronicity, we applied the 14-item Multitasking Preference Inventory (MPI) (Poposki & Oswald, 2010). Compared to other measures for polychronicity, this inventory measures the polychronicity on the individual level and not the cultural level. For measuring participants' social preferences, including their altruism, we relied on the Global Preference Survey (Falk, Becker, Dohmen, Enke, et al., 2018; Falk, Becker, Dohmen, Huffman, & Sunde, 2023). For altruism, participants had to answer two questions, one on a scale from 0 to 10 and one where they had to enter a value between \$0 and \$1600. A list with all items is attached in the Appendix (see Table D.5).

6.4.4.3. Manipulation Checks

The post-task questionnaire contained two manipulation check questions measured on a seven-point Likert scale. They tested whether instantiating features that allow the system to address the different preferences resulted in the desired effects. On a seven-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), participants were asked (1) whether "the microtask served an altruistic goal" to assess altruism, and (2) whether "[the participant] had to handle multiple tasks at once" to assess polychronicity.

6.4.4.4. Dependent Variables

Our main dependent variable is the job performance. We measured the job performance in two steps. First, we analyzed the number of provided alt-tags per participant and called this construct *quantity*. As all participants received the same payment for the alt-tag bonus task and all alt-tag submissions were completely voluntary, more provided alt-tags mean a higher job performance. For a deeper analysis, we also need to consider the quality (length and relevance) of the alt-tags. We define the construct $quantity_{adjusted}$ as follows:

$$\text{quantity}_{adjusted} = \text{quantity} \times \text{quality} \tag{6.1}$$

For each participant, we multiplied the quantity with a distinct quality factor. This quality factor for the participant is the mean of the quality score for all alt-tags that the participant provided and defined as follows:

$$\text{quality} = \frac{\sum (\text{relevance}_{normalized} \times 0.66 + \text{length}_{normalized} \times 0.33)}{\text{quantity}} \tag{6.2}$$

The quality of one alt-tag is defined by adding the relevance of the alt-tags and their length with different weights. We argue that both, the relevance and the length of the alt-tag are indicators of participants' job performance, but the relevance is twice as important as the length. A very short alt-tag, for example for a logo, might even be better than a very long alt-tag for the same logo. However, providing longer alt-tags shows more effort and also means that participants decided to also work on more complex images that require more text. For assessing the relevance of each alt-tag, we applied the scale of C. Williams et al. (2022) that provides four categories for alt-tags. As in their scale zero means that no alt-tag was provided, we do not need to include this category in our assessment. Consequently,

all alt-tags will receive scores from one to four depending on their relevance and specificity. To allow for an unbiased assessment, we developed a prompt and asked ChatGPT to assess each alt-tag. To do so, we provided ChatGPT the image, the Wikipedia page, the textual context of the image on the Wikipedia page, and a definition for each of the categories. Finally, we needed to normalize the resulting relevance score so that 0.25 means the alt-tag is of very low relevance for the image and one means that the alt-tag is very relevant and includes all necessary information. We did this by dividing the resulting scores by four.

To normalize the length of our alt-tags, we first counted the number of words for each alt-tag. We determined the 0.99 quantile of the maximum number of words for one alt-tag to exclude outliers. We then calculated the normalized length for each alt-tag by dividing the number of words by the number of words of the 0.99 quantiles.

We also measure additional dependent variables. First, we measured the Person-Job Fit on a seven-point Likert scale with the three items of Venkatesh et al. (2017). Second, we measured Job Satisfaction, also using a seven-point Likert scale. To do so, we took the three items of Sykes (2020). Third, for the Perceived Job Performance, we used a five-point Likert scale from *seldom* to *always* and took the items for task performance and context performance from Koopmans et al. (2014) as the other subconstructs did not apply to our context. For example, participants were not expected to participate in meetings, and there was no real organization that they could complain about. Consequently, these items were not useful for our task.

6.4.5. Recruitment

We recruited our participants on Prolific. Prolific is known to have many part-time crowdworkers for whom casual microtasking might be more convenient than for full-time crowdworkers (Oppenlaender, Milland, et al., 2020). For each of the two studies, we recruited 250 participants for the pre-screening. We recruited participants from the UK, US, and South Africa in similar portions to get a diverse set of altruistic values (Falk, Becker, Dohmen, Enke, et al., 2018; Falk, Becker, Dohmen, Huffman, & Sunde, 2023). We assumed that we would lose 20% of the 250 participants due to our exclusion criteria. We also excluded an additional 20% of participants who achieved average values for polychronicity (study 1) or altruism (study 2). We only included participants who had values lower or similar to the 40% quantile or similar or higher than the 60% quantile as we assumed that for average

crowdworkers our treatments would have no significant effects. Further, we assumed that we would lose another 20% of participants during or after the main task as they conducted too few main tasks, decided to not participate in the post-task questionnaire, or failed the attention checks in the post-task questionnaire. As we aimed for a minimum of 120 complete participants in each study, we decided to recruit 250 per study in the beginning. Participants received £1.20 for participating in the pre-screening. When successfully completing the main task, meaning submitting six of the seven tasks and having at least 50% of answers correct, they received a base payment of £6.50. They additionally received £0.10 for each correct answer in the main task as an incentive to answer the questions thoughtfully. For the alt-tag bonus tasks, they also received £0.10 per high-quality alt-tag. For participating in the post-task questionnaire, the crowdworkers received an additional bonus payment of £1.50.

6.4.6. Study 1: Polychronicity-Personalized System

To investigate the impact of a polychronicity-personalized casual microtasking system on crowdworkers' job performance, we conducted an experimental field study over seven consecutive days. In the study we had two different instantiations of our crowdsourcing system: the instantiation was either designed for polychrons or monochrons. Therefore, we had two treatment conditions. Participants used either a system that fitted their polychronicity or did not fit their polychronicity. We used LimeSurvey as the experimental platform in which we instantiated all questionnaires and provided access to install our casual microtaking system *MyCrowdSurfer*.

6.4.6.1. Pre-Screening

We invited 250 participants to the pre-screening for study 1. We had to exclude 43 participants due to our predefined criteria (attention checks, nationalities, primary browser, and English level). Additionally, we had to exclude one participant who did not enter a valid Prolific ID. We used the remaining 206 participants to standardize the mean polychronicity scores. Additionally, we added all single scores (Poposki & Oswald, 2010). Our participants covered almost the full range of possible answers for polychronicity (14 - 66) and showed with a mean of 39.12 and 36.41% of participants tending towards polychronicity (59.22% tending towards preferring monotasking) similar characteristics as other samples

(Lascău, Gould, Cox, et al., 2019). The resulting 40% quantile is -0.263, the 60% quantile is 0.377. Consequently, we characterize the 89 participants with polychronicity scores below or similar to -0.263 as monochrons and the 84 participants with altruism scores the same or higher than 0.377 as polychrons. We also controlled the sums of polychronicity scores so that we did not characterize participants with a tendency towards preferring monotasking as polychrons and the other way around. 33 participants were excluded from the next steps due to their polychronicity scores being not extreme enough. The remaining 173 participants were invited to the next step, the main task.

6.4.6.2. Sample characteristics

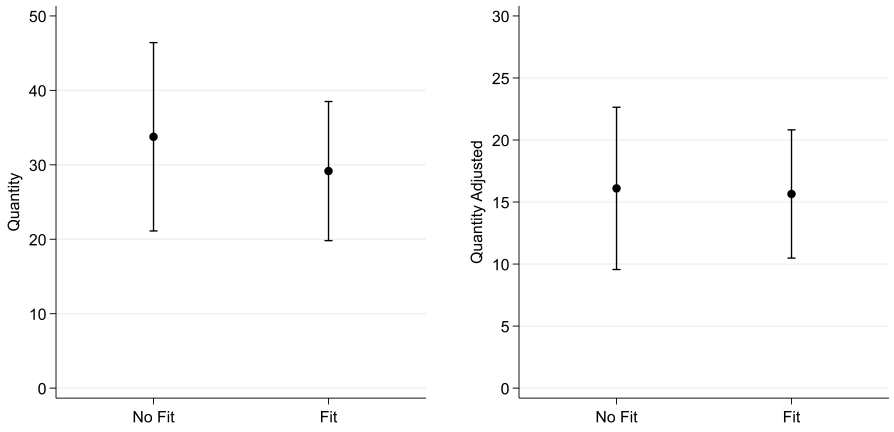
Out of the 173 participants invited to the main task, 95 participants did not fulfill the minimum requirements to successfully complete the second part of the study. These participants did not start the second task, did not complete the setup, or submitted less than six of the seven main tasks. This leaves us with responses from 78 participants for our analysis (51% female, $M_{age} = 35.6$), including 34 in the *No Fit* treatment and 44 in the *Fit* treatment. We provide more detailed sample characteristics about participants' demographics and individual preferences in Table D.6 and D.7 in the Appendix. Except for polychronicity, the demographics and preferences of participants did not differ statistically significantly between *Fit* and *No Fit* in study 1.

6.4.6.3. Manipulation Check

We conducted a manipulation check to evaluate the effectiveness of our two instantiations (multitasking and monotasking) in the polychronicity-personalized design. On a seven-point Likert scale, we asked crowdworkers whether they *"had to handle multiple tasks at once"*. Crowdworkers who interacted with the multitasking instantiation responded on average with 4.45, and crowdworkers who interacted with the monotasking instantiation responded on average with 3.82. Although this difference is not statistically significant (two-sided Mann-Whitney U test, $p = 0.159$), our experimental manipulation has the desired tendency. The non-significant differences might be due to limited statistical power.

6.4.6.4. Results

To test our Hypothesis 1a, we compare crowdworkers' performance in the bonus task (*quantity* and $quantity_{adjusted}$) depending on whether they interacted with a system that



(a) *Quantity* (b) $Quantity_{adjusted}$

Figure 6.5.: Study 1: Quantity and $Quantity_{adjusted}$ over Fit vs No Fit

fitted their individual polychronicity or not. We use an indicator variable *Fit*, which is one if crowdworkers interacted with a system that fitted their individual polychronicity and zero otherwise. Comparing the quantitative performance between *Fit* and *No Fit*, we find that the *quantity*, thus the average number of submitted alt-tags per participant, is 29.16 in the *Fit* treatment and 33.76 in the *No Fit* treatment (see Figure 6.5(a)). The difference is not statistically significant (two-sided Mann-Whitney U test, p = 0.746). When including the quality of the performance, taking into account the length and relevance of the alt tags, we find similar results. On average, $quantity_{adjusted}$ is 15.65 in the *Fit* treatment and 16.10 in the *No Fit* treatment (see Figure 6.5(b)). Again, the difference is not statistically significant (two-sided Mann-Whitney U test, p = 0.956).

Dep. Var.: Quantity	(1)	(2)	(3)	(4)
Fit	-4.606	-4.289	-2.594	-8.817
	(8.019)	(8.074)	(7.914)	(10.747)
Multitasking		9.867	9.549	0.814
		(8.070)	(7.928)	(13.692)
Fit x Multitasking				14.884
				(18.223)
Constant	33.765***	29.411***	49.789***	52.871***
	(6.442)	(7.625)	(17.400)	(18.572)
Demographics	✗	✗	✓	✓
R^2	0.005	0.025	0.056	0.067
Observations	78	78	76	76

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
For the complete table with all coefficients, see Table D.9 in the Appendix

Table 6.2.: Study 1: OLS regressions with *Quantity* as dependent variable

A series of OLS regressions[2] reported in Table 6.2 and Table 6.3 support the initial impression that there is no significant difference in performance between *Fit* and *No Fit*. To control whether the instantiation (see Figure 6.3) and the interaction between *Fit* and the instantiation affected the results, we used an indicator variable *Multitasking* which is one if crowdworkers interacted with the multitasking instantiation and zero if crowdworkers interacted with the monotasking instantiation. Model (1), (2), and (3) in Table 6.2 and Table 6.3 list the main effects of our *Fit* treatments and show that the coefficients are negative and not significant.

Dep. Var.: $Quantity_{adj.}$	(1)	(2)	(3)	(4)
Fit	-0.450	-0.280	0.668	-1.045
	(4.251)	(4.246)	(4.231)	(5.468)
Multitasking		5.305	4.726	2.322
		(4.350)	(4.203)	(7.305)
Fit x Multitasking				4.097
				(9.931)
Constant	16.100^{***}	13.759^{***}	21.025^{**}	21.873^{**}
	(3.330)	(3.627)	(9.169)	(9.698)
Demographics	✗	✗	✓	✓
R^2	0.000	0.021	0.045	0.048
Observations	78	78	76	76

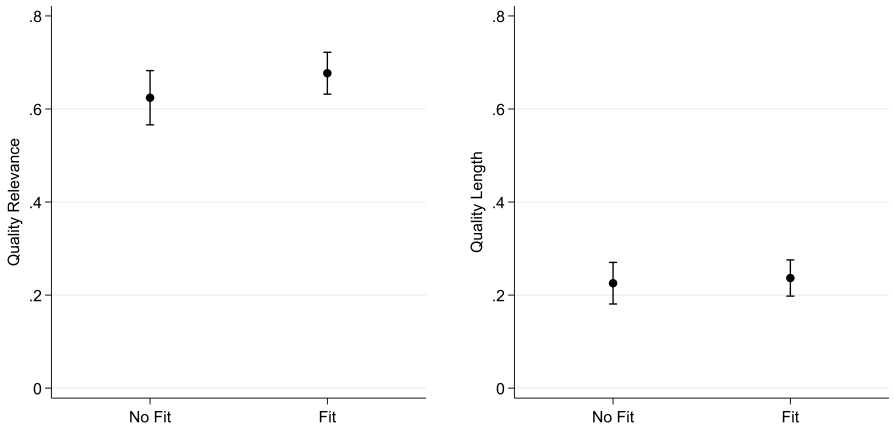
Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
For the complete table with all coefficients, see Table D.10 in the Appendix

Table 6.3.: Study 1: OLS regressions with $Quantity_{adjusted}$ as dependent variable

6.4.6.5. Additional Analyses

To better understand our results, we conducted additional analyses to assess the quality of crowdworkers' performance[3]. We analyzed the relevance of alt-tags and their length separately and compared them between *Fit* and *No Fit*. While we do not observe any large differences in the length of the alt-tags (see Figure 6.6(b)), we examined the relevance a little more closely (see Figure 6.6(a)). The average relevance of alt-tags is 0.677 in the *Fit* treatment and 0.624 in the *No Fit* treatment. Although this difference is not statistically significant (two-sided Mann-Whitney U test, p = 0.211), we do see that crowdworkers in the *Fit* treatment tend to provide better alt-tags with higher relevance on average. A series of OLS regressions confirm this tendency (see Table D.11 in the Appendix). The non-significant differences might be due to limited statistical power. We also report the

[2]In the models (3) and (4), two participants were not included because they answered *diverse* when asked about their gender.
[3]We used a subsample of 67 crowdworkers who provided at least one alt-tag for these additional analyses.



(a) *Quality Relevance* (b) *Quality Length*

Figure 6.6.: Study 1: Relevance and length of alt-tags over Fit vs No Fit

results for job satisfaction, person-job fit and fairness of payment in Appendix Table D.8.

6.4.7. Study 2: Altruism-Personalized System

6.4.7.1. Pre-Screening

We also invited 250 participants to the pre-screening for study 2. We had to exclude 40 participants due to our predefined criteria (attention checks, nationalities, primary browser, and English level). We used the remaining 210 participants to standardize and weigh the altruism scores according to Falk, Becker, Dohmen, Huffman, and Sunde (2023). Our participants covered almost the full range of possible answers for both altruism-related questions. The resulting 40% quantile is -0.0827, and the 60% quantile is 0.266. Consequently, we characterize the 84 participants with altruism scores below or similar to -0.0827 as selfish and the 87 participants with altruism scores the same or higher than 0.226 as altruistic. 39 participants were excluded from the next steps due to their altruism scores being not extreme enough. The remaining 171 participants were invited to the next step, the main task.

6.4.7.2. Sample characteristics

Out of the 171 participants invited to the main task of the second study, 103 participants did not fulfill the minimum requirements to successfully complete the second part of the study. These participants did not start the second task, did not complete the setup, or submitted less than six of the seven tasks. This leaves us with responses from 68 participants

for our analysis (51% female, $M_{age} = 33.2$), including 33 in the *No Fit* treatments and 35 in the *Fit* treatment. We provide more detailed sample characteristics about participants' demographics and individual preferences in Tables D.12 and D.13 in the Appendix. Similar to study 1, except for polychronicity, there are no statistically significant differences between *Fit* and *No Fit* in study 2.

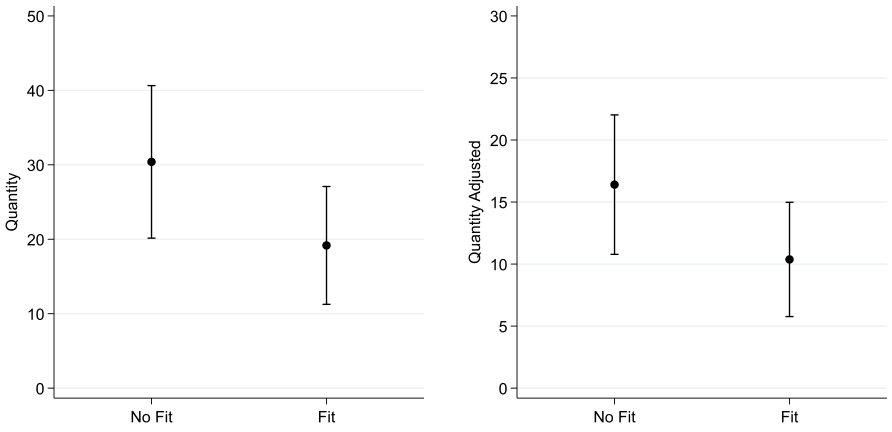
6.4.7.3. Manipulation Check

We conducted a manipulation check to evaluate the effectiveness of our two instantiations (selfish and altruistic) in the altruism-personalized design. On a seven-point Likert scale, we asked participants whether "the microtask served an altruistic goal". Crowdworkers who interacted with the altruistic instantiation responded on average with 4.73, and crowdworkers who interacted with the selfish instantiation responded on average with 3.71. This difference is statistically significant (two-sided Mann-Whitney U test, $p = 0.020$), and thus, our experimental manipulation performed as intended.

6.4.7.4. Results

To test Hypothesis 1b, we proceed similarly to the analysis of Hypothesis 1a. We compare crowdworkers' performance in the alt-tag bonus task (*quantity* and $quantity_{adjusted}$) depending on whether crowdworkers interacted with a system that fitted their individual altruistic preference or not. For the analysis in study 2, we use an indicator variable *Fit*, which is one if crowdworkers interacted with a system that fitted their individual altruistic preference and zero otherwise. The results show that *quantity*, thus the average number of alt-tags provided by each crowdworker, is 19.17 in the *Fit* treatment and 30.39 in the *No Fit* treatment (see Figure 6.7(a)). Thus, the results reveal that the quantitative performance was even higher in the *No Fit* treatment. However, this difference is not statistically significant (two-sided Mann-Whitney U test, $p = 0.191$). The results are similar for $quantity_{adjusted}$. On average, the $quantity_{adjusted}$ is 10.37 in the *Fit* treatment and 16.40 in the *No Fit* treatment (see Figure 6.7(b)). Again, the difference is not statistically significant (two-sided Mann-Whitney U test, $p = 0.166$).

We conducted a series of OLS regressions reported in Table 6.4 and Table 6.5 to further analyze the performance between *Fit* and *No Fit*. To control for the used instantiation, we used an indicator variable *Selfish*, which is one if crowdworkers interacted with the selfish



(a) *Quantity* (b) $Quantity_{adjusted}$

Figure 6.7.: Study 2: Quantity and $Quantity_{adjusted}$ over *Fit* vs *No Fit*

instantiation and zero if crowdworkers interacted with the altruistic instantiation. Model (1), (2), and (3) in Table 6.4 and Table 6.5 list the main effects of our *Fit* treatment. The coefficients are negative and, in some models, even marginally statistically significant. Thus, the results reveal that crowdworkers with a fit between their used instantiation and altruistic preference had no higher job performance.

Dep. Var.: Quantity	(1)	(2)	(3)	(4)
Fit	-11.223*	-10.609	-13.129*	-28.849***
	(6.605)	(6.688)	(6.876)	(9.293)
Selfish		6.680	3.045	-11.304
		(6.300)	(5.841)	(10.364)
Fit x Selfish				26.686*
				(13.769)
Constant	30.394***	26.345***	25.017*	27.191*
	(5.225)	(5.751)	(14.779)	(14.080)
Demographics	✗	✗	✓	✓
R^2	0.042	0.057	0.206	0.253
Observations	68	68	68	68

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
For the complete table with all coefficients, see Table D.15 in the Appendix

Table 6.4.: Study 2: OLS regressions with *Quantity* as dependent variable

6.4.7.5. Additional Analyses

The models (4) in the regression results in Tables 6.4 and 6.5 reveal that the coefficients for the interaction *Fit* x *Selfish* are at least marginally statistically significant. Thus, the results provide some indications that the crowdworkers' performance might differ depending on the selfish or altruistic instantiation. To understand the behavior in more

Dep. Var.: $Quantity_{adj.}$	(1)	(2)	(3)	(4)
Fit	-6.029	-5.674	-6.925*	-18.278***
	(3.707)	(3.817)	(3.870)	(5.217)
Selfish		3.869	1.754	-8.608
		(3.653)	(3.443)	(5.850)
Fit x Selfish				19.272**
				(7.408)
Constant	16.403***	14.058***	13.100	14.670
	(2.866)	(3.588)	(9.479)	(8.861)
Demographics	✗	✗	✓	✓
R^2	0.039	0.055	0.191	0.269
Observations	68	68	68	68

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
For the complete table with all coefficients, see Table D.16 in the Appendix

Table 6.5.: Study 2: OLS regressions with $Quantity_{adjusted}$ as dependent variable

detail, we again conducted additional analyses. Figures D.3 and D.4 in the Appendix provide an overview for *quantity* and $quantity_{adjusted}$ depending on whether we characterized crowdworkers as altruistic or selfish and whether they interacted with the altruistic or selfish instantiation. Contrary to our hypothesis, the results reveal that, on average, the altruistic crowdworkers in the altruistic instantiation (*Fit* treatment) had a lower performance compared to selfish crowdworkers in the altruistic instantiation and compared to altruistic and selfish crowdworkers in the selfish instantiation. First, we will analyze how the performance of altruistic crowdworkers differed between our two instantiations using $quantity_{adjusted}$ to also consider the quality of performance. $quantity_{adjusted}$ is on average 15.44 when altruistic crowdworkers interacted with the selfish instantiation and 5.46 when altruistic crowdworkers interacted with the altruistic instantiation. Although the performance almost tripled, a two-sided Mann-Whitney U test shows no statistically significant difference (p = 0.206). Again, we conducted a series of OLS regressions (see Table D.17 in the Appendix) and only considered altruistic crowdworkers. The coefficient for *Fit* is positive and statistically significant (model (1): p = 0.018 and model (2): p = 0.024) and reveals that altruistic crowdworkers had a significantly higher $quantity_{adjusted}$ when interacting with the selfish instantiation (*No Fit*) compared to the altruistic instantiation (*Fit*). Secondly, the results reveal that within the altruistic instantiation, altruistic crowdworkers had a statistically significant lower performance compared to selfish crowdworkers (two-sided Mann-Whitney U test, p = 0.028). $quantity_{adjusted}$ is, on average, 17.88 for selfish crowdworkers and 5.46 for altruistic crowdworkers. To analyze this impression with OLS regressions, we use an indicator variable *Altrusitic Pref.*, which is one if we character-

ized crowdworkers as altruistic and zero if we characterized crowdworkers as selfish. The main effects of *Altruistic Pref.* in the models (1) and (2) (see Table D.18) are positive and statistically significant (model (1): p = 0.011 and model (2): p = 0.018). We also report the results for job satisfaction, person-job fit and fairness of payment in Appendix Table D.14.

6.5. Discussion

Personalized task designs that adapt to crowdworkers' preferences are an innovative approach to impact crowdworkers' job performance. While research on personalized crowdworking in IS is rather scarce, there are promising results in other contexts like web personalization for affecting consumer behavior (S. Y. Ho & Bodoff, 2014; S. Y. Ho, Bodoff, & Tam, 2011; Zhang et al., 2011). Also, personalizing task assignments and recommendations in crowdworking and personalization according to cognitive abilities have been explored already. However, there is a research gap on personalization according to crowdworkers' preferences that go beyond their interests for specific topics or their favor for particular types of tasks. Against this backdrop, we first designed a personalized casual microtasking system following the DSR paradigm. We applied the P-E fit theory as our kernel theory to derive two requirements and justify our design instantiations. In the second step, we investigated the impact of preference-based personalization on job performance in casual microtasking systems in the context of collecting alt-tags on Wikipedia. In two field studies, we analyzed the impact of personalization according to the crowdworker's polychronicity (study 1) and altruism (study 2). In the following, we discuss the implications of our results for theory, design, and practice.

6.5.1. Theoretical Contributions

Our research first of all contributes to the descriptive knowledge base by establishing an evidence-based connection between personalized task designs and crowdworkers' job performance. With our longitudinal field study, we extend existing research that investigated the effect of P-J fit on job performance in traditional work settings as well as research on personalized crowdsourcing. Following P-J fit, we show in our first study, that polychronicity-based personalization does not increase the quantity of submitted bonus tasks. We assume this is caused by additional factors besides personalization that impact

job performance and wash out the effect of the personalization. For example, in the multitasking treatment, participants had more time to provide alt-tags and were more flexible and autonomous. They could submit alt-tags before, during, and after completing the main task while in the monotasking treatment, participants could only provide alt-tags after completing the main task. Also, some research suggests that the effect of polychronicity on job performance depends on the P-E fit (König & Waller, 2010), while other studies show that polychronicity, in general, is positively related to job performance (Kantrowitz et al., 2012). This could explain why we don't see a significant difference between participants in the *Fit* and *No Fit* treatment and polychrons providing slightly more alt-tags than monochrons (see Figure D.2 in the Appendix). Interestingly, the quality of alt-tags, especially the relevance, is higher for the *Fit* treatment than for the *No Fit* treatment. Therefore, our study provides evidence that polychronicity-based personalization can increase submission quality. This could be explained by monotaskers being able to focus on one task at a time and therefore providing more relevant and detailed alt-tags. At the same time, multitaskers benefit from providing alt-tags directly in the context. This effect needs to be investigated in further studies.

In our second study, we show that contrary to our hypotheses, personalization according to altruism leads to fewer submitted bonus tasks. Also, the quality is not better when a fit between the crowdworkers' social preference and the system design exists. This is contrary to the idea of P-J fit. Interestingly, while in the selfish system design, both groups performed similarly, altruistic crowdworkers' provided significantly fewer alt-tags when they were using the altruistic design. Our manipulation checks show that our designs work and altruistic participants thought that the altruistic design served more an altruistic goal than the selfish design. Although the altruistic goal was provided externally by us as the task requester, we assume that altruistic workers internalized the altruistic motive (referred to as internalized extrinsic motivation (Ryan & Deci, 2000)) and were, therefore, both intrinsically and extrinsically motivated to provide alt-tags. However, the interplay of intrinsic and extrinsic motivators (in our case bonus payments) is very complex. The introduction of monetary incentives has the potential to strengthen, but also to reduce intrinsic motivation, as explained by the Motivation Crowding Theory (Frey & Jegen, 2001). Monetary incentives can be counterproductive for prosocial activities due to the image motivation being crowded out. In a public scenario, individuals would not have

the feeling of being perceived positively by others for their prosocial behavior due to the additional financial incentive (Ariely et al., 2009). Also, they are concerned about seeming to be greedy instead of prosocial (Exley, 2017). Monetary benefits for prosocial behavior can also create doubt about the true motive of the altruistic activity and make individuals lose their intrinsic motivation. This is called the *overjustification effect* (Bénabou & Tirole, 2006). Although the core driver for the overjustification effect is the social reputation that can only suffer in a public scenario, the effect can also happen in a private setting where individuals then start questioning their own motives, leading to a reduced intrinsic motivation (Bénabou & Tirole, 2006). This effect is also related to the *tainted altruism* phenomenon that refers to situations where altruistic acts are perceived as having selfish motivations (Newman & Cain, 2014). In our case, the monetary reward makes the act of providing alt-tags less pure as it is mixed with self-interests. Additionally, as shown by Cassar and Meier (2021) our participants could have also seen the prosocial incentive as a strategic move by us rather than a genuine act of kindness.

We assume that a combination of these phenomena explains the reduced job performance, in terms of quantity and quality of alt-tags, of altruistic crowdworkers who participated in the altruistic design. Selfish crowdworkers did not internalize the altruistic goal of the task and therefore were still mainly extrinsically motivated which shows in a similar task performance as in the selfish treatment. Also, participants might have questioned our altruistic motives as the task requester as the monetary incentive for alt-tags made it seem like highlighting the altruistic goal could also be a strategic move. An interesting aspect of our study is that contrary to most existing studies that experience crowding out effects, we compare a pure extrinsic motivation in the selfish design with a combination of additional intrinsic motivation for altruistic crowdworkers in the altruistic design. The effect that we identified also contradicts research on crowdworkers' motivations to participate in crowdworking tasks, which often includes aspects like having an impact (Deng & Joshi, 2016). Based on our results, we can not determine if this effect would happen for all kinds of social preferences or is specific to altruism. This is worth investigating in further studies.

6.5.2. Design Contributions

We contribute with two requirements and an artifact instantiation to the prescriptive knowledge base. We, therefore, complement prior research endeavors to improve job per-

formance in crowdsourcing tasks based on personalization (Pagano & Bruegge, 2013). Despite our results showing unexpected effects, our study provides prescriptive knowledge that can serve as a solid foundation to inform the design of further personalized casual microtasking systems. We rigorously analyzed the design instantiations in a longitudinal field study to demonstrate the feasibility of personalized casual microtasking. We also show that our system is applicable in the context of crowdsourcing image alt-tags on Wikipedia and demonstrate the advantages of casual microtasking in a new context.

Our two theory-driven design requirements exploiting the P-E fit theory provide context-specific design knowledge to guide the design of preference-based personalization in casual microtasking. We show that personalization according to polychronicity (REQ1) does not improve overall job performance but can lead to better quality. In our design instantiation, we show how the design of casual microtasking can also respect the preferences of monochrons, not forcing them into a multitasking behavior as it would happen in a traditional task design (Haug, Benke, Fischer, & Maedche, 2023). Regarding personalization according to social preferences (REQ2), we provide evidence that altruism-based personalization can lead to contrary effects as suggested by P-J fit, especially for altruistic preferences. We contribute a design instantiation that demonstrates how crowdworking task designs can be adapted to different social preferences. In our case, we showed how to make the goal of a task seem more altruistic to crowdworkers.

Both requirements provide high generalizability to other crowdworking tasks. While the polychronicity-based personalization of the timing of feedback tasks is rather specific to casual microtasking, personalized task recommendations according to polychronicity scores for other, more traditional crowdworking tasks, might be beneficial for the quality of results. While some crowdworking tasks require participants to switch between tabs, like verifying information, others require more focus, like transcriptions. Consequently, some tasks might fit better to polychrons while others fit the preferences of monochrons.

For our second requirement, we focused on altruism as a social preference as this fit the context of our task. Of course, the requirement could also be implemented by addressing other social preferences or task types. However, based on our results we recommend being cautious when combining intrinsic and extrinsic motivators in crowdworking. In accordance with existing research (M.-H. Wu & Quinn, 2017), we demonstrate that changing the wording in task instructions and task management can have an impact on the task

outcomes. Therefore, task requesters need to be careful when designing their tasks. Although in our study, both system designs would overall lead to similar results, we show that it is beneficial to know the social preferences of participants in casual microtasking.

6.5.3. Practical Contributions

With our study, we support practitioners with actionable knowledge to improve distinct aspects of job performance in crowdworking. Low-quality crowdworking results are a persistent problem of crowdworking platforms. While past research often focused on quality measures or the effects of varying financial incentives in crowdwork (Daniel et al., 2018; C. J. Ho et al., 2015), we followed another approach, accounting for and leveraging the individual differences of crowdworkers. Also, existing research on personalized crowdsourcing mainly focused on task assignments, not adapting the tasks themselves. Therefore, our research makes two practical contributions to personalized crowdworking.

First, our proposed design guides how to drive task quality and quantity by utilizing different personalization mechanisms. Practitioners and researchers can use this design knowledge for their task design. They can build upon our results to make a funded decision for or against preference-based personalization in their tasks. Moreover, they can also use the design knowledge to inform their task design even if deciding against personalization. Our first design instantiations demonstrate how to instantiate crowdworking tasks that require multitasking or monotasking behavior. However, our practical contribution is not limited to the proposed instantiations only. Our task designs could be transferred to real use cases as well. While in our case, we introduced an artificial main task to only allow monotasking, in a real-world scenario, mouse movements could be used to detect when monotaskers are open to work on a task as suggested in previous work (Paulino, Correia, Barroso, & Paredes, 2023). A simpler implementation would be a button with which casual microtasks could be requested. As in related research, users preferred customizable user interfaces compared to static or adaptive ones (Eichler & Dostál, 2012), allowing users to choose between a monotasking and multitasking design themselves could also be beneficial. This would also allow users to adapt the task designs to their current context or mood. Our second design instantiation shows practitioners how to display or hide the goal of a task. We contribute three general components of crowdworking tasks that can be adapted to display different motives. This design can easily be transferred to all kinds of

crowdworking tasks. However, as our results show, personalization according to altruism needs to be treated carefully. It is beneficial to understand the social preferences of the participants to avoid negative effects due to the effect of tainted altruism (Newman & Cain, 2014). More generally, our study shows that the interplay of intrinsic and extrinsic motivators in crowdwork is complex and that it is not always better to show crowdworkers that they can make an impact with their submissions. Moreover, related research also raised concerns regarding the ethical implications of introducing intrinsic motivation to crowdworking tasks (Law et al., 2016). Additional intrinsic motivations might lead to crowdworkers unknowingly contributing more without getting paid more. As our study shows, additional intrinsic motivation is not always beneficial and does not always lead to crowdworkers contributing more. However, when task requesters still decide to add an intrinsic motivator, they need to ensure that payments are still fair.

Second, we contribute to practice with two artifacts that provide exemplary instantiations for two different types of preference-based personalization of casual microtasking. These artifacts exemplify how practitioners and researchers can apply the proposed design knowledge in real-world crowdworking contexts. We demonstrate the application of these systems in two real-world casual microtasking scenarios using Prolific. In doing so, we address potential reasons for the low job performance of crowdworkers in casual microtasking. This enables practitioners to successfully build further personalized casual microtasking systems and learn from our results. Therefore, our research is highly relevant to practice in that it enables a better understanding of personalized task designs in crowdworking as an approach to increase the job performance of crowdworkers.

6.5.4. Limitations and Future Research Opportunities

Our work is not without limitations that also provide opportunities for future research. First, for our study, we needed to introduce an artificial main task to replace a real task. This main task was necessary to be able to distinguish between a monotasking and multitasking design by knowing when participants are finished with their primary task without having to rely on behavioral data like click data that is difficult to obtain in a field study. Future research could investigate features that allow for a monotasking design in a real-world scenario.

Second, we only investigated personalization according to one social preference, namely

altruism. While this is a well-researched preference that also fits our task context, there are of course more social preferences and we can not necessarily transfer our results to all other social preferences. Each social preference might have specific effects and needs to be addressed in different ways. This means that the results in our study, namely the effect of tainted altruism can not necessarily be generalized to other social preferences. Consequently, our study only serves as a starting point to investigate more social preferences in future studies.

Third, we expect a self-selection bias in our studies. We explained to participants from the beginning that they would need to install a Google Chrome extension to participate in the task to minimize costly dropouts. Participants who were not willing to do so as they do not prefer such kinds of tasks or have data privacy concerns probably refrained from participation. There is no way in research to completely avoid self-selection bias. Also, in a real-world scenario, we would experience the same bias as crowdworkers who did not want to install browser extensions would not participate in casual microtasks on Prolific. Therefore, we argue that our results are still generalizable to all potential users of casual microtasking systems.

Fourth, we conducted a rigorous field study, using an innovative artifact. We decided on a field study, to guarantee the empirical validity of our results. Although the artifact was extensively tested before the study, there might have been minor technical issues. As we expect these issues to not be specific to one instantiation or one group of participants, we suppose they would not affect our results. Future research could investigate the detected effects in a more controlled setting like a lab study. This would allow for a deeper understanding of the robustness and generalizability of our results.

6.6. Conclusion

Our research addresses the important challenge of improving the job performance of crowdworkers in casual microtasking. We propose a theory-driven design for preference-based personalization of casual microtasking by building on the P-E fit theory. We instantiated the two derived requirements in innovative software artifacts. To understand the effect of personalization on job performance of crowdworkers we conducted two experimental field studies on Prolific. We show that personalization can have a positive impact on the quality of submitted tasks when personalizing the task design according to crowdworkers'

polychronicity. Further, we demonstrate that the effect of personalization according to crowdworkers' altruism as a social preference can have detrimental effects on job performance. We explain this by the overjustification effect and tainted altruism when combining an extrinsic motivator with an altruistic motive. Our study contributes to the prescriptive knowledge base with two theoretically grounded design requirements. Additionally, we contribute to the descriptive knowledge base by providing insights into the complex effects of preference-based personalization on job performance in casual microtasking.

7. Discussion

Crowd-feedback systems enable the scalable evaluation of interactive and static designs. While existing research has demonstrated the feasibility of asking a large group of people for design feedback and has shown that the outcomes can achieve a similar quality as expert feedback (Yuan et al., 2016), research lacks a deeper understanding of the individual design features and design decisions that make crowd-feedback systems successful. Following a human-centered design approach focusing on crowdworkers' needs and requirements is promising to provide solutioins addressing these challenges. Understanding the effects of specific design features on crowdworker and the resulting feedback can also enable the development of design recommendations for tailoring crowd-feedback systems to specific use cases. However, a major problem of crowd feedback persists. Crowdworkers are not in a real context of use when providing feedback and are not actual or potential users of the software which could affect their feedback negatively. Therefore, further research is required on how to combine traditional user feedback in the form of questionnaires and pop-ups on websites and the contemporary crowd feedback approach.

In this thesis, I explore the design of human-centered crowd-feedback systems. Specifically I inverstigate the two design challenges of human-centered crowd-feedback systems, namely, how to tailor crowd-feedback systems to specific use cases and goals by understanding the effects of design features and how to combine crowd feedback and traditional user feedback to put crowdworkers in a real context of use when providing feedback. To address these design challenges, I designed, developed, and evaluated four systems. The results of these studies have several theoretical contributions and practical implications, which I will discuss in the following. Subsequently, I will discuss the major limitations of these studies and propose promising future work that addresses these limitations and go beyond the insights derived from the previous studies.

7.1. Theoretical Contributions

The five studies of this thesis make several theoretical contributions that are summarized in Table 7.1. First, I will present the theoretical contributions for each study individually and how they address their associated research questions. Subsequently, I will summarize

Table 7.1.: Summary of the Theoretical Contributions of this Dissertation.

Study	Theoretical Contributions
Study I	• Conceptualization of crowd feedback that can describe any crowd feedback approach along the dimension input, crowd configuration, design characteristics and effects • Identification of research gaps and three core research streams as guidance for future research
Study II	• Identification of the most important design features of crowd-feedback systems from a crowdworker's perspective • Prescriptive knowledge in the form of three design implications that explain how to design crowd-feedback system for achieving defined goals
Study III	• Prescriptive knowledge in the form of four design rationales for designing configuration systems for crowd-feedback systems • Prescriptive knowledge in the form of a configuration process for the design of individual crowd-feedback systems
Study IV	• Prescriptive knowledge in the form of three design rationales for a crowd-feedback system that integrates tasks into everyday internet surfing • Six design recommendations for browser extensions to integrate crowdsourcing tasks in everyday internet surfing • A generalizable design pattern regarding the design of casual microtasking systems in the form of browser extensions
Study V	• Prescriptive knowledge in the form of two design requirements for designing personalized crowdsourcing systems • Descriptive knowledge in the form of a deeper understanding of how personalization impacts crowdworkers' job performance

the overall theoretical contributions of this thesis and explain how they address the RQs of this dissertation.

Study I provides a systematic literature review of crowd-feedback systems. Research on crowd-feedback systems is spread over various domains like IS and computer science, but most papers are found in the HCI domain. The research topic is scattered and multiple different terms are used that can all be summarized under the term "crowd-feedback system". I demonstrate the rise of crowd feedback and related systems and the consequent need for a systematic overview.

Based on the established methodology, I rigorously conducted an SLR (Webster & Watson, 2002) and coded the 40 identified papers according to Nickerson et al. (2013) and the grounded-theory approach by Wolfswinkel et al. (2013). Thereby, I identified four dimensions and 28 characteristics. Based on Morschheuser et al. (2017), Pedersen et al. (2013), and Zuchowski et al. (2016), I developed a conceptual framework for crowd feedback and identified four dimensions of crowd feedback. I created a morphological box (Zwicky & Wilson, 1967) that is structured along the main dimensions of Input, Crowdsourcing Configuration, Design Dimensions, and Effects. Therefore, I provide an answer to sub-RQ1a: *how to conceptualize crowd feedback for IS development?*

To answer RQ1b of *what is the state-of-the-art of crowd-feedback in IS development and what are future research directions?*, I created a concept matrix that provides a structured overview of the characteristics of existing research. Based on the concept matrix, I identify three research streams by conducting a cluster analysis that also helps to better understand the current state-of-the-art. Based on the concept matrix and existing research streams, I identified four avenues for future research.

This comprehensive approach contributes by allowing for the description of any crowd-feedback system along the dimensions and characteristics identified. The literature-grounded systematic overview provided by the conceptualization and the morphological box presents a key contribution of this study. Further, study I contributes by identifying and explaining three research streams of crowd-feedback systems that help to classify existing systems but also allow to structure research in this field in the future. The study can also serve as a starting point for researchers and practitioners by also highlighting existing research gaps and offering ideas for future research avenues. This work is fundamental for the de-

velopment of a shared understanding of crowd feedback and is therefore the basis for my dissertation.

In **study II**, I first investigate the design of an interactive crowd-feedback system to answer the question of *how to design a human-centered crowd-feedback system to evaluate interactive designs? (RQ2a)*. Particularly, I designed the crowd-feedback system *Feeasy* based on an interview study with ten participants. Based on the interviews, I could derive five core design features for crowd-feedback systems for evaluating interactive designs and designed *Feeasy*, an interactive crowd-feedback system. This was then followed by an experimental study to understand the effects of the five core design features of *Feeasy* and answer the question of *how do different crowd-feedback system design features affect crowdworkers' perceptions and the resulting feedback quality and quantity? (RQ2b)*. For the second part, I draw on the TIME (Sundar, Jia, et al., 2017) to inform the study design. I assumed that different design features impact the perceived interactivity of a crowd-feedback system, which in turn has an effect on the user engagement, and the user's behavior in terms of feedback quality and quantity. In the experimental study, I had seven treatments, one individual for each design feature, one treatment with no design features included, and one treatment with all five design features combined. The experimental study was followed by an interview study with 28 participants to better understand crowdworker's perceptions of the features.

This study first contributes by identifying and evaluating the five core design features of crowd-feedback systems from a crowdworker's perspective. Although the quantitative results did not show significant differences between the treatments, the qualitative insights helped to understand how the features and their combination impacts the perceptions and behavior of crowdworkers. There seems to be an information overload effect (Roetzel, 2019) for crowdworkers. They seemed to be overwhelmed when the crowd-feedback system offered all five features. They first need to get familiar with every single one before starting to provide feedback. Also, some features lead to more specific and focused feedback, like markers, while others helped to share more generic feedback, like the star rating. By summarizing the results, this study also contributes by deriving three implications for the design of crowd-feedback systems for specific goals, like optimizing for crowdworkers' perceptions, feedback quality, or feedback quantity.

In **study III**, I addressed the need to make crowd-feedback systems available for designers

with no development skills and knowledge in crowdsourcing design feedback. The goal of the study was to answer the question of *how to design a configuration system to support designers in creating effective customized crowd-feedback requests? (RQ3)*. Therefore, I relied on existing knowledge of configuration systems (cf. Randall et al., 2007), end-user development (cf. Lieberman et al., 2006), and conducted an exploratory literature review on crowd-feedback systems and related processes. Additionally, I conducted expert interviews (N=14) that show that despite the crowd-feedback systems can address multiple persistent problems of evaluation processes, they are never applied in practice. To address this challenge, I designed a parameter-based configuration system for *Feeasy* and conducted a focus group evaluation with 10 participants.

This study contributes four design rationales for designing a configuration system for crowd-feedback systems highlighting the importance of user guidance, system customization, explanation of effects of design decisions, and understanding the crowdworkers' perspective. Related research in the field of end-user development, configuration systems, or tools for supporting design processes could build upon these design rationales. Further, I contribute a configuration process to design and adapt crowd-feedback systems. Finally, the study offers a summarization of experts' opinions on the configuration system. For example, I discovered a need for balancing the flexibility and complexity of the configuration system.

In **study IV**, I investigated the integration of crowd-feedback tasks into the daily internet surfing of crowdworkers. Thereby, I address the question of *how to design a system to collect in situ crowd feedback in the form of casual microtasking to improve the working conditions of crowdworkers and feedback quality? (RQ4)*. To answer this question, I build upon the work of (Hahn et al., 2019; Seyff, Graf, & Maiden, 2010; Seyff, Ollmann, & Bortenschlager, 2014) on the integration of microtasks in general or feedback tasks in particular in other activities. To develop design rationales, I conducted semi-structured interviews with five crowdworkers. Based on the design rationales, I developed the *CrowdSurfer* system. In the final evaluation, I conducted a field study with 63 participants, in which I compared the feedback that is collected via the *CrowdSurfer* with feedback that is collected via a traditional survey. Finally, I conducted interviews with 12 of the participants in the *CrowdSurfer* treatment regarding their perceptions of the usability, their process to provide feedback, and the impact of the *CrowdSurfer* on their working

conditions and motivation.

Therefore, this study offers three theoretical contributions: First, the study provides design rationales that were derived from interviews with crowdworkers about the design of microtasks that can be integrated in crowdworkers' daily internet surfing. To the best of my knowledge, this is the first study, that investigates the design of a Chrome extension for integrating microtasks in daily internet surfing and therefore offers unique insights. Second, I generalized the result of the field study and provided six design recommendations for the design of such browser extensions. Third, I provide a generalizable pattern for such casual microtasking systems that can be transferred to other types of tasks.

In **study V**, I address the research gap of preference-based personalized crowdworking. In the study, I first address the question of *how to design a preference-based personalized casual microtasking system to increase job performance? (RQ5a)*. To answer the question, I develop two theory-driven design requirements exploiting the P-E fit theory (Caplan, 1987; Edwards, Caplan, & Van Harrison, 1998). Based on the requirements and the *CrowdSurfer* design of study IV, I develop two instantiations of the *MyCrowdSurfer* system a preference-based personalized casual microtasking system for the collection of alt-tags on Wikipedia. In the second step, I aim to answer the question of *how do preference-based personalizations in casual microtasking systems affect job performance? (RQ5b)*. I conducted two longitudinal field studies using the instantiations of the *MyCrowdSurfer* system to answer the question.

Consequently, this study contributes descriptive knowledge in the form of two theory-driven design requirements guiding the design of preference-based personalized casual microtasking systems. These requirements can be easily transferred to other crowdworking task types. Moreover, I contribute prescriptive knowledge by analyzing the effect of personalization on job performance. I show in this study, that polychronicity-based personalization has the potential to increase crowdworkers' job performance. Regarding altruism-based personalization, I contribute insights into the complex relationship between intrinsic and extrinsic motivators and provide evidence that altruism-based personalization has detrimental effects on altruistic crowdworkers.

In summary, all five studies provide prescriptive knowledge to shape research on designing human-centered crowd-feedback systems to make evaluation processes more scalable and

at the same time deliver high-quality feedback outcomes. With study I this thesis provides a solid theoretical foundation for the following studies. study I also provides a literature-driven conceptualization and overview of crowd-feedback systems that help to cluster and structure design features. Further, this study allowed me to identify research gaps in previous research on crowd feedback. This is important since human-centered and scalable evaluation of IS is crucial for IS success and user acceptance and crowd feedback has the potential to solve persistent issues of traditional methods.

In studies II, IV, and V, I focused on understanding design features and their effects from the perspective of feedback providers in order to give recommendations for the design of crowd-feedback systems for specific use cases. This is important, first, because understanding the effects is important for creating effective crowd-feedback systems and optimizing the feedback outcomes. Second, both studies allowed for a more seamless integration of crowd feedback into the interaction with the IS, design, or prototype. This also impacts the perceptions of feedback providers and is important to gather real and relevant in situ feedback. Study III provides additional design knowledge from a feedback requester perspective and thereby extends the results of study II. This is important as the design of a configuration system can bridge the gap between research on crowd feedback and the actual application of designers in practice. Thereby, the developed artifacts serve as instantiations for the contributed design knowledge of this dissertation. The experimental evaluations with crowdworkers and design experts in the form of field studies, interviews, and focus groups provide a profound understanding of how feedback processes look and how the design impacts the behavior of feedback providers. Consequently, this dissertation supports researchers with design knowledge of crowd-feedback systems and an understanding of integrating feedback tasks into a natural interaction which informs future designs of crowdsourcing systems for feedback collection.

Beyond these findings, this thesis contributes design knowledge for crowdsourcing systems in general. Study IV investigates the design of a casual microtasking system in the form of a browser extension. The developed design knowledge also contributes to crowdsourcing research in general and can be the foundation of the development of further tools, processes, and patterns for integrating crowdsourcing tasks into other primary activities. Also, study V contributes to crowdsourcing research in general by showcasing the effect of adapting the microtask design to crowdworkers' preferences (i.e., polychronicity and altruism). The

gained insights can be used for optimizing existing or future crowdsourcing systems and can improve the results of crowdwork for practitioners. Also, the insights can serve as a starting point for further investigations on adaptive crowdsourcing systems to improve job performance.

7.2. Practical Contributions

Table 7.2.: Summary of the Practical Implications of this Dissertation.

Study	Practical Implications
Study I	• Overview of existing work on crowd-feedback systems along a framework including input factors, design characteristics of crowd-feedback systems, crowd characteristics, and effects of applying the systems
Study II	• Design instantiation of an interactive crowd-feedback system • Classification of design features to provide or enrich design feedback • An understanding of crowdworkers' perceptions of these different design features
Study III	• Design instantiation of a configuration system for creating and tailoring crowd-feedback requests • An understanding of practitioners' needs when creating and tailoring crowd-feedback requests
Study IV	• Design instantiation of a casual microtasking system for crowdsourcing design feedback • Demonstration of the potential of the integration of microtasks into crowdworkers' everyday internet surfing • Delivering a potential solution for improving crowdworkers' working conditions
Study V	• Design instantiation of a preference-personalized casual microtasking system for crowdsourcing alt-tags for images

For practitioners, this dissertation contributes with system artifacts, applicable knowledge, and empirical evaluations which provide important practical implications. I summarize the practical implications in Table 7.2.

With **study I**, I contribute foundational knowledge on crowd-feedback systems. The study

informs practitioners about research on crowd feedback in general, and crowd-feedback systems and their features in more particular. Designers and developers who seek to collect design feedback themselves could use this study to learn about different design features of crowd-feedback requests and how they are instantiated in different systems. They can choose from the presented features which features they want to implement in their system and learn about combinations and effects of these features according to the presented papers. Also, my conceptualization can help them to find crowd-feedback systems and related papers for their use case to gain an even deeper understanding.

In **study II** and **study III**, two developed design artifacts serve as exemplary instantiation of a crowd-feedback system for static and interactive designs that can be configured by designers themselves. Designers who want to apply crowd feedback in practice but struggle to develop crowd-feedback systems themselves as they lack development knowledge can use the configuration system to build their own instantiation of *Feeasy*. Moreover, study II informs about the effects of design features on users' perceptions, such as lower ease of use when combining too many features. Study III provides insights into the requirements of designers when using such configuration systems. Therefore, developers who aim to build configuration systems for similar use cases can build upon these insights.

In **study IV**, I contribute another design artifact in the form of a browser extension, that instantiates the developed design rationales. This innovative design instantiation allows the integration of crowdsourcing tasks into the everyday internet surfing of crowdworkers. Further, I demonstrate in this study, that casual microtasking is feasible for crowdsourcing and provide an overview of the perspective of crowdworkers regarding this crowdsourcing approach which can help in the design of further casual mirotasking systems. The *Crowd-Surfer* system provides a flexible design that can be adapted to many other microtasks and can therefore serve as a basis for the development of further casual microtasking systems. Finally, I contribute a process for casual microtasking that can be leveraged by practitioners to build business models around casual microtasking. This process includes the creation and handling of microtasks as well as managing the payments.

Finally, in **study V**, I contribute a preference-personalized casual microtasking system, *MyCrowdSurfer*, that is built upon the artifact of study IV. This artifact instantiates two requirements that were derived from the P-E fit theory in the context of crowdsourcing alt-tags on Wikipedia. I also provide recommendations for practitioners on how and where

to consider social preferences in the design of microtasks in general. In the study, I investigated a system that specifically personalizes to crowdworkers' polychronicity and altruism. However, the design can be generalized and the system can be adapted to many other preferences. The implementation of the *MyCrowdSurfer* system may therefore also serve as a template.

Concluding, I want to summarize the implications for practitioners, including designers, developers of feedback systems, and requesters on crowdsourcing platforms. My dissertation offers a comprehensive overview of potential inputs, design features and effects of design features on feedback providers and the feedback itself. I contribute four artifacts that instantiate design rationales in the context of crowd feedback and can serve as templates for the application of crowd feedback and casual microtasking in practice.

7.3. Limitations and Future Research

Although I have rigorously followed established scientific methodology to plan, conduct, and report all five studies, they still come with limitations. Beyond, the explicitly mentioned limitations in the respective studies, I will explain the overarching limitations of this dissertation in this section and highlight future research directions that evolve from them.

First, I need to point out, that my dissertation only focused on the feedback collection process, neglecting the subsequent steps like analyzing the resulting feedback and deriving implications for the design. While it is of course beneficial to crowdsource feedback to get many diverse opinions, this also increases the effort for analyzing the resulting feedback. The same goes for preferring qualitative over quantitative feedback. On the one hand qualitative feedback contains more information as feedback providers can share their thoughts including reasons, ideas and problems, on the other hand, analyzing qualitative feedback can be tedious work and is hard to automate. Although there are approaches that handle the analysis of large amounts of feedback like *Feedbackmap* (Beeferman & Gillani, 2023) or *Decipher* (Yen, Kim, & Bailey, 2020), these are limited to a specific use case and can not necessarily handle the results of all feedback features in my studies. Covering the analysis of feedback would make the artifacts more comprehensive and applicable in practice. Also, being able to summarize the resulting feedback and visualize it could also allow for the exploration of a new dimension of the feedback quality.

Second, rigorously measuring the quality of feedback is difficult. In study II and study IV, I relied on established categories for describing the quality of design feedback, like specificity and relevance. While asking crowdworkers to assess the quality of all collected feedback comments ensures objectivity, it also risks that crowdworkers might not fully understand the importance of some feedback comments as they are not the designer of the website that is evaluated. Novices might also interpret design comments much differently than expert designers (Foong et al., 2017). I tried to counteract this problem by ensuring that the crowdworkers who evaluate the feedback have experience in UI/UX design. Still, they did not know about the intentions and design goals of the original designers of the websites and could therefore not consider this in their evaluation. In study V, I applied a different approach and relied on the capabilities of GPT4 to evaluate the provided alt-tags according to predefined categories. Although this approach might also have its drawbacks, it is less biased by the personal opinions of individuals. I assume a combination of different feedback assessment techniques, like crowdworkers, experts, and automated means like GPT4 would lead to the best and most objective results. However, due to time and cost restrictions, this was not implementable for my studies.

Third, in all of my studies, I used an artificial context for the feedback collection. It was clear to participants in my evaluation studies that they participate in research studies. That might have influenced the motivation of crowdworkers to participate in the studies and provide feedback. I assume that in some studies the feedback would have been different if the feedback requester had been the website owner or designer herself. However, as the artificial context was the same for all participants in all treatments, the identified effects of designs are still valid and can not be attributed to the artificial context of feedback collection.

Related to this, I purely focused my studies on the evaluation of website designs. Also, the four artifacts were built for website designs and cannot necessarily be used for other types of information systems. Especially, the *CrowdSurfer* can only be used for collecting feedback on website designs. However, I am confident that the contributed design knowledge is generalizable to the evaluation of other types of information systems. Still, future research needs to investigate real use cases in practice, for example by conducting field studies together with industry partners to guarantee the empirical validity and generalizability of my results.

Fifth, the feedback features and their instantiation in study II could have influenced the results of this study and study III. While the features were picked according to an initial interview study, their instantiation was arbitrary, considering related work that used the same features. For example, the category features can be implemented in many different ways. Categories could be implemented by allowing users to add their comments to design principles (Yuan et al., 2016) or offering the possibility to share different types of feedback like a first notice, impressions, and opinions on the fulfillment of design goals (Xu & Bailey, 2014). Using different categories or implementing them in another way could have also influenced the behavior of users of *Feeasy*. As we used the same features in study III, the perceptions of the participants in the focus group workshops could also depend on the instantiation that I decided for.

Sixth, my studies were limited to a short period and did not analyze the longitudinal effects of the instantiated systems. While studies II and III were limited to a one-time usage of the system, studies IV and V ran for seven days. Still, this period might be too short to understand how casual microtasking affects crowdworkers' working behavior. Also, the effects of personalization might vary when crowdworkers use the system over a longer period. Especially, the effect of casual microtasking on work-life boundaries could be worth investigating. In the context of crowd feedback it could be interesting to understand how the provided feedback will change the more experienced crowdworkers get in providing design feedback.

Lastly, recently generative artificial intelligence in the form of large language models (LLMs) has become popular in many areas of practice and research. Of course, LLMs could also be helpful in creating, improving, or analyzing design feedback. In my studies, I did not include any LLM for two reasons: First, when I started with my dissertation and planned my studies, LLMs were not available yet. Second, my dissertation shall mainly provide a foundation for crowd-feedback research as there was a lack of knowledge for crowd-feedback systems and therefore I did focus on the core elements in the first step. However, I believe that there are useful ways of applying LLMs in the context of crowd feedback, for example, to help feedback providers to come up with creative suggestions, give hints to feedback providers on how to improve the feedback or summarize the feedback as it is for example done in the *Feedbackmap* already (Beeferman & Gillani, 2023).

8. Conclusion

User involvement in IS development processes is an important step when aiming for high user satisfaction and overall development success. Traditional methods like usability tests and focus groups suffer from scalability and cost issues and are not applicable in every step of the development process. An emerging but underexplored approach is crowdsourcing feedback, meaning that a large undefined group of people are asked for feedback. For feedback collection, dedicated crowd-feedback systems can be used that offer various functionalities to collect high-quality design feedback.

This dissertation deals with understanding human-centered crowd-feedback systems in general, analyzing their design, and investigating approaches to improve feedback outcomes. The studies presented in this dissertation aim to bridge the gap between theoretical knowledge and practical implementation to contribute to the body of knowledge on crowd feedback as a feasible method to scale IS evaluation. Particularly, I tackle two main research gaps. First, my dissertation aims at understanding and leveraging the design features of crowd-feedback systems. Second, I address the integration of crowd feedback with traditional user feedback mechanisms to place crowdworkers in more realistic contexts of use when providing feedback. This thesis contributes to five studies that investigate these design challenges and provide answers to the related research questions. In study I, I provide a comprehensive overview of the state-of-the-art of crowd feedback. I also contribute a conceptualization of crowd-feedback systems, identifying key input factors, design characteristics, crowd configurations, and effects. Further, I identified three core research streams of crowd-feedback systems and derived interesting avenues for future research. My following studies build upon the results of study I. In studies II and III, I built a crowd-feedback system from a human-centered perspective. After understanding the effects of five core design features of crowd-feedback systems on crowdworkers' perceptions and feedback outcomes, I used this design knowledge to build a configuration system that enables novice designers to build and adapt crowd-feedback systems according to their context and desired outcomes. In studies IV and V, I built two artifacts that allow the integration of microtasks into crowdworkers' everyday internet surfing. The benefit of this so-called casual microtasking is that crowdworkers are in a context of use when providing feedback on a website, while also getting paid for their feedback. In study IV, I focused on analyzing the feasibility of this approach and its impacts on crowdworkers' perceptions

and job performance: I conducted an experimental online study, followed by qualitative interviews to get a deeper understanding. In study V, I investigated the effects of preference-based personalization in casual microtasking. To improve the job performance of crowdworkers, I integrated a preference-adaptive design in the artifact of the fourth study to analyze the effect of personalization in an experimental field study. Thereby, I contribute with descriptive and prescriptive knowledge on the design of personalized casual microtasking systems, especially in the context of feedback tasks. In conclusion, this dissertation makes a significant contribution to the field of IS evaluation by offering both theoretical insights and practical implementations. Therefore, this thesis helps to make IS evaluation processes more scalable and human-centered to enable designers and developers to successfully develop human-centered IS.

Bibliography

Abbas, T., & Gadiraju, U. (2022). Goal-Setting Behavior of Workers on Crowdsourcing Platforms: An Exploratory Study on MTurk and Prolific. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *10*(1), 2–13. https://doi.org/10.1609/HCOMP.V10I1.21983

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected papers of hirotugu akaike* (pp. 199–213). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0{_}15

Almaliki, M., Ncube, C., & Ali, R. (2014). The Design of Adaptive Acquisition of Users Feedback: An Empirical Study. *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, 1–12. https://doi.org/10.1109/RCIS.2014.6861076

Alpar, P., & Osterbrink, L. (2018). Antecedents of Perceived Fairness in Pay for Microtask Crowdwork. *26th European Conference on Information Systems, ECIS 2018*, 1–13.

Alsayasneh, M., Amer-Yahia, S., Gaussier, E., Leroy, V., Pilourdault, J., Borromeo, R. M., Toyama, M., & Renders, J. .-. (2018). Personalized and Diverse Task Composition in Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, *30*(1), 128–141. https://doi.org/10.1109/TKDE.2017.2755660

Alyahya, S. (2020). Crowdsourced Software Testing: A Systematic Literature Review. *Information and Software Technology*, *127*. https://doi.org/10.1016/j.infsof.2020.106363

Ambreen, T., & Ikram, N. (2016). A state-of-the-art of empirical literature of crowdsourcing in computing. *Proceedings - 11th IEEE International Conference on Global Software Engineering, ICGSE 2016*, 189–190. https://doi.org/10.1109/ICGSE.2016.37

Amer-Yahia, S., Gaussier, E., Leroy, V., Pilourdault, J., Borromeo, R. M., & Toyama, M. (2016). Task Composition in Crowdsourcing. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 194–203. https://doi.org/10.1109/DSAA.2016.27

Andreoni, J., Harbaugh, W. T., & Vesterlund, L. (2010). Altruism in Experiments. *Behavioural and Experimental Economics*, 6–13. https://api.semanticscholar.org/CorpusID:15793245

Anish, P. R., & Ghaisas, S. (2014). Product Knowledge Configurator for Requirements Gap Analysis and Customizations. *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 437–443. https://doi.org/10.1109/RE.2014.6912295

Ariely, D., Bracha, A., & Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, *99*(1), 544–555. https://doi.org/10.1257/aer.99.1.544

Asghar, M., Gull, N., Tayyab, M., Zhijie, S., & Tao, X. (2020). Polychronicity at work: Work engagement as a mediator of the relationships between job outcomes. *Journal of Hospitality and Tourism Management*, *45*, 470–478. https://doi.org/10.1016/J.JHTM.2020.10.002

Asghar, M., Tayyab, M., Gull, N., Zhijie, S., Shi, R., & Tao, X. (2021). Polychronicity, work engagement, and turnover intention: The moderating role of perceived organizational support in the hotel industry. *Journal of Hospitality and Tourism Management*, *49*, 129–139. https://doi.org/10.1016/j.jhtm.2021.09.004

Awad, N. F., & Krishnan, M. S. (2006). The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization. *MIS Quarterly*, *30*(1), 13–28. https://doi.org/10.2307/25148715

Ayalon, O., & Toch, E. (2018). Crowdsourcing Privacy Design Critique: An Empirical Evaluation of Framing Effects. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2018-Janua*, 4752–4761. https://doi.org/10.24251/hicss.2018.598

Ayalon, O., & Toch, E. (2019). A/P(rivacy) Testing: Assessing Applications for Social and Institutional Privacy. *Extended Abstracts of the 2019 CHI Conference*, 1–6. https://doi.org/10.1145/3290607.3312972

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: Technological Antecedents and Implications. *MIS Quarterly*, *35*(4), 831–858. https://doi.org/10.2307/41409963

Bandara, W., Furtmueller, E., Gorbacheva, E., Miskon, S., & Beekhuyzen, J. (2015). Achieving rigor in literature reviews: Insights from qualitative data analysis and tool-support. *Communications of the Association for Information Systems*, *37*(1), 154–204. https://doi.org/10.17705/1cais.03708

Beeferman, D., & Gillani, N. (2023). FeedbackMap: A Tool for Making Sense of Open-ended Survey Responses. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 395–397. https://doi.org/10.1145/3584931.3607496

Bénabou, R., & Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, *96*(5), 1652–1678. https://doi.org/10.1257/aer.96.5.1652

Berg, J. (2016). Income Security in the On-Demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers. *Comparative Labor Law and Policy Journal*, *37*(3), 27. https://api.semanticscholar.org/CorpusID:146876380

Bluedorn, A. C., Kalliath, T. J., Strube, M. J., & Martin, G. D. (1999). Polychronicity and the Inventory of Polychronic Values (IPV):The development of an instrument to measure a fundamental dimension of organizational culture. *Journal of Managerial Psychology*, *14*(3/4), 205–231. https://doi.org/10.1108/02683949910263747/FULL/PDF

Bodendorf, F., & Kaiser, C. (2009). Detecting opinion leaders and trends in online social networks. *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining*, 65. https://doi.org/10.1145/1651437.1651448

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brhel, M., Meth, H., Maedche, A., & Werder, K. (2015). Exploring Principles of User-Centered Agile Software Development: A Literature Review. *Information and Software Technology*, *61*, 163–181. https://www.sciencedirect.com/science/article/pii/S0950584915000129

Buettner, R. (2015). A Systematic Literature Review of Crowdsourcing Research from a Human Resource Management Perspective. *2015 48th Hawaii International Conference on System Sciences*, 4609–4618. https://doi.org/10.1109/HICSS.2015.549

Cable, D. M., & Edwards, J. R. (2004). Complementary and supplementary fit: A theoretical and empirical integration. *Journal of Applied Psychology*, *89*(5), 822–834. https://doi.org/10.1037/0021-9010.89.5.822

Cambridge University Press & Assessment. (2023). Personalization. https://dictionary.cambridge.org/us/dictionary/english/personalization

Caplan, R. D. (1987). Person-environment fit theory and organizations: Commensurate dimensions, time perspectives, and mechanisms. *Journal of Vocational Behavior*, *31*(3), 248–267. https://doi.org/https://doi.org/10.1016/0001-8791(87)90042-X

Cassar, L. (2018). Job Mission as a Substitute for Monetary Incentives: Benefits and Limits. *Management Science*, *65*(2), 896–912. https://doi.org/10.1287/mnsc.2017.2903

Cassar, L., & Meier, S. (2021). Intentions for Doing Good Matter for Doing Well: The Negative Effects of Prosocial Incentives. *The Economic Journal*, *131*(637), 1988–2017. https://doi.org/10.1093/ej/ueaa136

Charness, G., Cobo-Reyes, R., & Sánchez, A. (2016). The effect of charitable giving on workers' performance: Experimental evidence. *Journal of Economic Behavior & Organization*, *131*, 61–74. https://doi.org/https://doi.org/10.1016/j.jebo.2016.08.009

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. http://www.jstor.org/stable/4132490

Chen, C. Y., Yen, C.-H., & Tsai, F. C. (2014). Job crafting and job engagement: The mediating role of person-job fit. *International Journal of Hospitality Management*, *37*, 21–28. https://doi.org/https://doi.org/10.1016/j.ijhm.2013.10.006

Chilton, M. A., Hardgrave, B. C., & Armstrong, D. J. (2005). Person-Job Cognitive Style Fit for Software Developers: The Effect on Strain and Performance. *Journal of Management Information Systems*, *22*(2), 193–226. http://www.jstor.org/stable/40398750

Chiou, W. C., Lin, C. C., & Perng, C. (2010). A Strategic Framework for Website Evaluation based on a Review of the Literature from 1995-2006. *Information and Management*, *47*(5-6), 282–290. https://doi.org/10.1016/j.im.2010.06.002

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining / KDD-2001*, 263–268. https://doi.org/10.1145/502512.502549

Choi, Y., Monserrat, T.-J. J. K. P., Park, J., Shin, H., Lee, N., & Kim, J. (2021). ProtoChat: Supporting the Conversation Design Process with Crowd Feedback. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *4*(CSCW3), 19–23. https://doi.org/10.1145/3432924

Codagnone, C., Abadie, F., & Biagi, F. (2017). The Future of Work in the Sharing Economyy. Market Efficiency and Equitable Opportunities or Unfair Precarisation? *SSRN Electronic Journal*, *JRC101280*, 96. https://doi.org/10.2139/ssrn.2784774

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys*, *51*(1), 1–40. https://doi.org/10.1145/3148148

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly: Management Information Systems*, *13*(3), 319–339. https://doi.org/10.2307/249008

de la Cruz, G. V., Peng, B., Lasecki, W. S., & Taylor, M. E. (2015). Towards Integrating Real-Time Crowd Advice with Reinforcement Learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, 17–20. https://doi.org/10.1145/2732158.2732180

DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2022). Estimating Social Preferences and Gift Exchange at Work. *American Economic Review*, *112*(3), 1038–1074. https://doi.org/10.1257/aer.20190920

DellaVigna, S., & Pope, D. (2018). What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies*, *85*(2), 1029–1069. https://doi.org/10.1093/restud/rdx033

Deng, X., & Joshi, K. D. (2016). Why Individuals Participate in Micro-task Crowdsourcing Work Environment: Revealing Crowdworkers' Perceptions. *Journal of the Association for Information Systems*, *17*, 648–673. https://doi.org/10.17705/1jais.00441

Deng, X., Joshi, K. D., & Galliers, R. D. (2016). The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful Through Value Sensitive Design. *MIS Quarterly: Management Information Systems*, *40*(2), 279–302. https://doi.org/10.25300/MISQ/2016/40.2.01

Detert, J. R., Treviño, L. K., & Sweitzer, V. L. (2008). Moral Disengagement in Ethical Decision Making: A Study of Antecedents and Outcomes. *Journal of Applied Psychology*, *93*(2), 374–391. https://doi.org/10.1037/0021-9010.93.2.374

Difallah, D. E., Demartini, G., & Cudré-Mauroux, P. (2013). Pick-a-Crowd: Tell Me What You like, and I'll Tell You What to Do. *WWW '13: Proceedings of the 22nd international conference on World Wide Web*. https://doi.org/10.1145/2488388.2488421

Dobler, D., Friedrich, S., & Pauly, M. (2020). Nonparametric MANOVA in Meaningful Effects. *Annals of the Institute of Statistical Mathematics*, *72*(4), 997–1022. https://doi.org/10.1007/s10463-019-00717-3

Dow, S., Gerber, E., & Wong, A. (2013). A Pilot Study of Using Crowds in the Classroom. *Proceedings of the 2013 Conference on Human Factors in Computing Systems*, 227–236. https://doi.org/10.1145/2470654.2470686

Durward, D., Blohm, I., & Leimeister, J. M. (2016). Crowd Work. *Business & Information Systems Engineering*, *58*(4), 281–286. https://doi.org/10.1007/s12599-016-0438-0

Easterday, M. W., Rees Lewis, D., & Gerber, E. M. (2017). Designing Crowdcritique Systems for Formative Feedback. *International Journal of Artificial Intelligence in Education*, *27*(3), 623–663. https://doi.org/10.1007/s40593-016-0125-9

Ebert, N., Scheppler, B., Ackermann, K. A., & Geppert, T. (2023). QButterfly: Lightweight Survey Extension for Online User Interaction Studies for Non-Tech-Savvy Researchers. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3544548.3580780

Edwards, J. R., Caplan, R. D., & Van Harrison, R. (1998). Person-Environment Fit Theory: Conceptual Foundations, Empirical Evidence, and Directions for Future Research. In *Theories of organizational stress* (pp. 28–67). Oxford University Press. https://doi.org/10.1093/oso/9780198522799.003.0003

Edwards, J. R., & Shipp, A. J. (2007). The Relationship between Person-Environment Fit and Outcomes: An Integrative Theoretical Framework. *Perspectives on Organizational Fit*, 209–258. https://api.semanticscholar.org/CorpusID:146289804

Eichler, Z., & Dostál, M. (2012). Adaptive user interface personalization in ERP systems. *International Conference on Business Information Systems*, 49–60.

Exley, C. (2017). Incentives for Prosocial Behavior: The Role of Reputations. *Management Science*, *64*(5), 2460–2471. https://doi.org/10.1287/mnsc.2016.2685

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences. *The Quarterly Journal of Economics*, *133*(4), 1645–1692. https://doi.org/10.1093/QJE/QJY013

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. *Management Science*, *69*(4), 1935–1950. https://doi.org/10.1287/MNSC.2022.4455

Fan, H., & Poole, M. (2006). What Is Personalization? Perspectives on the Design and Implementation of Personalization in Information Systems. *Journal of Organizational Computing and Electronic Commerce*, *16*, 179–202. https://doi.org/10.1207/s15327744joce1603{\&}4{_}2

Fehr, E., & Fischbacher, U. (2002). Why Social Preferences Matter - The Impact of Non-Selfish Motives on Competition, Cooperation and Incentives. *The Economic Journal*, *112*(478), C1–C33. http://www.jstor.org/stable/798356

Fehr, E., & Schmidt, K. M. (2001). Theories of fairness and reciprocity-evidence and economic applications. *Available at SSRN 264344*.

Feine, J., Morana, S., & Maedche, A. (2019). Designing a Chatbot Social Cue Configuration System. *Proceedings of the 40th International Conference on Information Systems (ICIS)*, 17. https://aisel.aisnet.org/icis2019/design_science/design_science/2

Feine, J., Morana, S., & Maedche, A. (2020). Designing Interactive Chatbot Development Systems. *ICIS 2020 Proceedings*. https://aisel.aisnet.org/icis2020/is_workplace_fow/is_workplace_fow/9

Felstiner, A. (2011). Working the Crowd: Employment and Labor Law in the Crowdsourcing Industry. *Berkeley Journal of Employment and Labor Law*, *32*(1), 19–24. https://doi.org/10.3109/9780203490891-6

Foong, E., Gergle, D., & Gerber, E. M. (2017). Novice and expert sensemaking of crowdsourced feedback. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–18. https://doi.org/10.1145/3134680

Frey, B. S., & Jegen, R. (2001). Motivation Crowding Theory. *Journal of Economic Surveys*, *15*(5), 589–611. https://doi.org/10.1111/1467-6419.00150

Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007). MyExperience: A system for In Situ Tracing and Capturing of User Feedback on Mobile Phones. *MobiSys'07: Proceedings of the 5th International Conference on Mobile Systems, Applications and Services*, 57–70. https://doi.org/10.1145/1247660.1247670

Gadiraju, U., Kawase, R., & Dietze, S. (2014). A Taxonomy of Microtasks on the Web. *HT '14: Proceedings of the 25th ACM conference on Hypertext and social media*, 218–223. https://doi.org/10.1145/2631775.2631819

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, 1631–1640. https://doi.org/10.1145/2702123.2702443

Geiger, D., & Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems — Current state of the art. *Decision Support Systems*, *65*, 3–16. https://doi.org/https://doi.org/10.1016/j.dss.2014.05.007

Gibbs, A. (1997). Focus Groups. *Social Research Update*, *19*(8), 1–8.

Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., & Kostakos, V. (2013). Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. *UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 753–762. https://doi.org/10.1145/2493432.2493481

Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E., & Kostakos, V. (2015). Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Computer Networks*, *90*, 34–48. https://doi.org/https://doi.org/10.1016/j.comnet.2015.07.002

Grady, C., & Lease, M. (2010). Crowdsourcing Document Relevance Assessment with Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating*

Speech and Language Data with Amazon's Mechanical Turk, 172–179. https://aclanthology.org/W10-0727

Greenberg, M. D., Easterday, M. W., & Gerber, E. M. (2015). Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. *C and C 2015 - Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 235–244. https://doi.org/10.1145/2757226.2757249

Grimm, P. (2010, December). Social Desirability Bias. In *Wiley international encyclopedia of marketing*. https://doi.org/10.1002/9781444316568.wiem02057

Guan, Y., Deng, H., Risavy, S. D., Bond, M. H., & Li, F. (2011). Supplementary Fit, Complementary Fit, and Work-Related Outcomes: The Role of Self-Construal. *Applied Psychology*, *60*(2), 286–310. https://doi.org/10.1111/J.1464-0597.2010.00436.X

Guay, F., Vallerand, R., & Blanchard, C. (2000). On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion*, *24*, 175–213. https://doi.org/10.1023/A:1005614228250

Hahn, N., Iqbal, S. T., & Teevan, J. (2019). Casual Microtasking: Embedding Microtasks in Facebook. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300249

Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial Least Squares Structural Equation Modeling (PLS-SEM): An Emerging Tool in Business Research. *European Business Review*, *26*(2), 106–121. https://doi.org/10.1108/EBR-10-2013-0128

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574

Hara, K., Milland, K., Hanrahan, B. V., Callison-Burch, C., Adams, A., Savage, S., & Bigham, J. P. (2019). Worker Demographics and Earnings on Amazon Mechanical Turk: An Exploratory Analysis. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312970

Harris, M. A., & Weistroffer, H. R. (2009). A new look at the relationship between user involvement in systems development and system success. *Communications of the Association for Information Systems*, *24*(1), 739–756. https://doi.org/10.17705/1cais.02442

Haug, S., Benke, I., Fischer, D., & Maedche, A. (2023). CrowdSurfer: Seamlessly Integrating Crowd-Feedback Tasks into Everyday Internet Surfing. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3544548.3580994

Haug, S., Benke, I., Fischer, D., Walther, S., Nieken, P., & Maedche, A. (2023). *Preference-based Personalization of Casual Microtasking for Crowdworkers.*

Haug, S., Benke, I., & Maedche, A. (2023). Aligning Crowdworker Perspectives and Feedback Outcomes in Crowd-Feedback System Design. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–28. https://doi.org/10.1145/3579456

Haug, S., & Maedche, A. (2021a). Crowd-Feedback in Information Systems Development: A State-of-the-Art Review. *Proceedings of the 42nd International Conference on Information Systems (ICIS) 2021*, 1–17. https://aisel.aisnet.org/icis2021/is_design/is_design/4

Haug, S., & Maedche, A. (2021b). Feeasy: An Interactive Crowd-Feedback System. *Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology, UIST 2021*, 41–43. https://doi.org/10.1145/3474349.3480224

Haug, S., Sommerrock, S., Benke, I., & Maedche, A. (2023). Scalable Design Evaluation for Everyone! Designing Configuration Systems for Crowd-Feedback Request Generation. *Proceedings of Mensch und Computer 2023 (MuC 2023)*, 91–100. https://doi.org/10.1145/3603555.3603566

Haukipuro, L., Pakanen, M., & Väinämö, S. (2016). Online User Community for Efficient Citizen Participation. *Proceedings of the 20th International Academic Mindtrek Conference*, 78–85. https://doi.org/10.1145/2994310.2994341

Hecht, T. D., & Allen, N. J. (2005). Exploring links between polychronicity and well-being from the perspective of person–job fit: Does it matter if you prefer to do only one thing at a time? *Organizational Behavior and Human Decision Processes*, *98*(2), 155–178. https://doi.org/https://doi.org/10.1016/j.obhdp.2005.07.004

Hetmank, L. (2013). Components and Functions of Crowdsourcing Systems – A Systematic Literature Review. *Wirtschaftsinformatik Proceedings 2013*, 55–69. https://doi.org/10.13140/2.1.3836.4166

Hettiachchi, D., van Berkel, N., Hosio, S., Kostakos, V., & Goncalves, J. (2019). Effect of Cognitive Abilities on Crowdsourcing Task Performance. *Human-Computer Interaction – INTERACT 2019*, 442–464.

Hettiachchi, D., van Berkel, N., Kostakos, V., & Goncalves, J. (2020). CrowdCog: A Cognitive Skill Based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–22. https://doi.org/10.1145/3415181

Hettiachchi, D., Wijenayake, S., Hosio, S., Kostakos, V., & Goncalves, J. (2020). How Context Influences Cross-Device Task Acceptance in Crowd Work. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 53–62. https://doi.org/10.1609/hcomp.v8i1.7463

Hicks, C. M., Pandey, V., Fraser, C. A., & Klemmer, S. (2016). Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 458–469. https://doi.org/10.1145/2858036.2858195

Hiltz, S. R., & Turoff, M. (1985). Structuring Computer-Mediated Communication Systems to Avoid Information Overload. *Communications of the ACM*, *28*(7), 680–689. https://doi.org/10.1145/3894.3895

Ho, C. J., Slivkins, A., Suri, S., & Vaughan, J. W. (2015). Incentivizing High Quality Crowdwork. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 419–429. https://doi.org/10.1145/2736277.2741102

Ho, S. Y., & Bodoff, D. (2014). The Effects of Web Personalization on User Attitude and Behavior. *MIS Quarterly*, *38*(2), 497–A10. https://www.jstor.org/stable/26634936

Ho, S. Y., Bodoff, D., & Tam, K. Y. (2011). Timing of Adaptive Web Personalization and Its Effects on Online Consumer Behavior. *Information Systems Research*, *22*(3), 660–679. http://www.jstor.org/stable/23015600

Hosseini, M., Shahri, A., Phalp, K., & Ali, R. (2016). Crowdsourcing Transparency Requirements through Structured Feedback and Social Adaptation. *Proceedings -*

International Conference on Research Challenges in Information Science. https://doi.org/10.1109/RCIS.2016.7549330

Howard, M. C., & Cogswell, J. E. (2023). A meta-analysis of polychronicity: Applying modern perspectives of multitasking and person-environment fit. *Organizational Psychology Review*, *13*(3), 315–347. https://doi.org/10.1177/20413866221143370

Howe, J. (2006). The Rise of Crowdsourcing. *Wired magazine*, *14*(6), 1–4. https://www.wired.com/2006/06/crowds/

Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business* (1st ed.). Crown Publishing Group.

Hsu, J. S., Hung, Y. W., Chen, Y.-H., & Huang, H.-H. (2013). Antecedents and Consequences of User Coproduction in Information System Development Projects. *Project Management Journal*, *44*(2), 67–87. https://doi.org/10.1002/pmj.21330

Huang, N., Burtch, G., Gu, B., Hong, Y., Liang, C., Wang, K., Fu, D., & Yang, B. (2018). Motivating User-Generated Content with Performance Feedback: Evidence from Randomized Field Experiments. *Management Science*, *65*(1), 327–345. https://doi.org/10.1287/mnsc.2017.2944

Hui, J. S., Greenberg, M. D., & Gerber, E. M. (2014). Understanding the Role of Community in Crowdfunding Work. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 62–74. https://doi.org/10.1145/2531602.2531715

Huxhold, O., Li, S. C., Schmiedek, F., & Lindenberger, U. (2006). Dual-Tasking Postural Control: Aging and the Effects of Cognitive Demand in Conjunction with Focus of Attention. *Brain Research Bulletin*, *69*(3), 294–305. https://doi.org/10.1016/j.brainresbull.2006.01.002

Imas, A. (2014). Working for the "warm glow": On the benefits and limits of prosocial incentives. *Journal of Public Economics*, *114*, 14–18. https://doi.org/https://doi.org/10.1016/j.jpubeco.2013.11.006

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–620. https://doi.org/10.1145/2470654.2470742

ISO 9241-210. (2019). Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. https://www.iso.org/standard/77520.html

Ives, B., & Olson, M. H. (1984). User Involvement and MIS Success: A Review of Research. *Management Science*, *30*(5), 586–603. https://doi.org/10.1287/mnsc.30.5.586

Jäger, G., Zilian, L. S., Hofer, C., & Füllsack, M. (2019). Crowdworking: working with or against the crowd? *Journal of Economic Interaction and Coordination*, *14*(4), 761–788. https://doi.org/10.1007/s11403-019-00266-1

Jansson, A. D., & Bremdal, B. A. (2018). Genetic Algorithm for Adaptable Design using Crowdsourced Learning as Fitness Measure. *2018 International Conference on Smart Systems and Technologies (SST)*, 1–6. https://doi.org/10.1109/SST.2018.8564686

Kang, H. B., Amoako, G., Sengupta, N., & Dow, S. P. (2018). Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174180

Kantrowitz, T. M., Grelle, D. M., Beaty, J. C., & Wolf, M. B. (2012). Time is money: Polychronicity as a predictor of performance across job levels. *Human Performance*, *25*(2), 114–137. https://doi.org/10.1080/08959285.2012.658926

Kaplan, T., Saito, S., Hara, K., & Bigham, J. (2018). Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 70–78. https://doi.org/10.1609/hcomp.v6i1.13327

Kasunic, A., Chiang, C. W., Kaufman, G., & Savage, S. (2019). Turker Tales: Integrating Tangential Play into Crowd Work. *DIS 2019 - Proceedings of the 2019 ACM Designing Interactive Systems Conference*, 21–34. https://doi.org/10.1145/3322276.3322359

Kaufman, C. F., Lane, P. M., & Lindquist, J. D. (1991). Exploring More than 24 Hours a Day: A Preliminary Investigation of Polychronic Time Use. *Journal of Consumer Research*, *18*(3), 392. https://doi.org/10.1086/209268

Kaufman-Scarborough, C., & Lindquist, J. D. (1999). Time management and polychronicity. *Journal of Managerial Psychology*, *14*(3/4), 288–312. https://doi.org/10.1108/02683949910263819

Kirchberg, D. M., Roe, R. A., & Van Eerde, W. (2015). Polychronicity and Multitasking: A Diary Study at Work. *Human Performance*, *28*(2), 112–136. https://doi.org/10.1080/08959285.2014.976706

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.471

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The Future of Crowd Work. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*, 1301–1318. https://doi.org/10.1145/2441776.2441923

Knaeble, M., Nadj, M., & Maedche, A. (2020). Oracle or Teacher? A Systematic Overview of Research on Interactive Labeling for Machine Learning. In *Wirtschaftsinformatik proceedings 2020* (pp. 2–16). GITO Verlag. https://doi.org/10.30844/wi{_}2020{_}a1-knaeble

Koch, I., Gade, M., Schuch, S., & Philipp, A. M. (2010). The role of inhibition in task switching: A review. *Psychonomic Bulletin & Review*, *17*(1), 1–14. https://doi.org/10.3758/PBR.17.1.1

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking-An integrative review of dual-task and task-switching research. *Psychological bulletin*, *144*(6), 557–583. https://doi.org/10.1037/BUL0000144

Kokoç, M. (2019). Flexibility in e-Learning: Modelling Its Relation to Behavioural Engagement and Academic Performance. https://api.semanticscholar.org/CorpusID:219148492

König, C. J., & Waller, M. J. (2010). Time for Reflection: A Critical Examination of Polychronicity. *Human Performance*, *23*, 173–190. https://api.semanticscholar.org/CorpusID:144787022

Koopmans, L., Bernaards, C., Hildebrandt, V., Buuren, S., van der Beek, A., & De Vet, H. (2014). Improving the Individual Work Performance Questionnaire using Rasch Analysis. *Journal of applied measurement*, *15*, 160–175. https://doi.org/10.1136/oemed-2013-101717.51

Kossek, E. E., Ruderman, M. N., Braddy, P. W., & Hannum, K. M. (2012). Work–nonwork boundary management profiles: A person-centered approach. *Journal of Vocational Behavior*, *81*(1), 112–128. https://doi.org/https://doi.org/10.1016/j.jvb.2012.04.003

Krause, M., Garncarz, T., Song, J., Gerber, E. M., Bailey, B. P., & Dow, S. P. (2017). Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4627–4639. https://doi.org/10.1145/3025453.3025883

Kreiss, E., Bennett, C. L., Hooshmand, S., Zelikman, E., Morris, M. R., & Potts, C. (2022). Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:248987078

Kristof, A. L. (1996). Person-Organization Fit: An Integrative Review of its Conceptualizations, Measurement, and Implications. *Personnel Psychology*, *49*(1), 1–49. https://doi.org/10.1111/j.1744-6570.1996.tb01790.x

Kurup, A. R., & Sajeev, G. P. (2018). Task Personalization for Inexpertise Workers in Incentive Based Crowdsourcing Platforms. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 286–292. https://doi.org/10.1109/ICACCI.2018.8554475

Lakhani, K. R., Boudreau, K. J., Loh, P. R., Backstrom, L., Baldwin, C., Lonstein, E., Lydon, M., MacCormack, A., Arnaout, R. A., & Guinan, E. C. (2013). Prize-based contests can provide solutions to computational biology problems. *Nature biotechnology*, *31*(2), 108–111. https://doi.org/10.1038/nbt.2495

Lakhani, K. R., Garvin, D. A., Lonstein, E., R. Lakhani, K., Garvin, D. A., & Lonstein, E. (2010). TopCoder (A) Developing Software through Crowdsourcing. *Harvard Business School Case*, (610-032), 1–22. https://papers.ssrn.com/abstract=2002884

Lascău, L., Gould, S. J. J., Brumby, D. P., & Cox, A. L. (2022). Crowdworkers' Temporal Flexibility is Being Traded for the Convenience of Requesters Through 19 'Invisible Mechanisms' Employed by Crowdworking Platforms. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3491101.3519629

Lascău, L., Gould, S. J. J., Cox, A. L., Karmannaya, E., & Brumby, D. P. (2019). Monotasking or Multitasking: Designing for Crowdworkers' Preferences. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300649

Law, E., Yin, M., Goh, J., Chen, K., Terry, M., & Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. *Conference on Human Factors in Computing Systems - Proceedings*, 4098–4110. https://doi.org/10.1145/2858036.2858144

Leicht, N. (2018). Given Enough Eyeballs, all Bugs are Shallow - A Literature Review for the Use of Crowdsourcing in Software Testing. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 4102–4111. https://doi.org/10.24251/HICSS.2018.515

Lekschas, F., Ampanavos, S., Siangliulue, P., Pfister, H., & Gajos, K. Z. (2021). Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 12. https://doi.org/10.1145/3411764.3445507

Leung, A. C. M., Santhanam, R., Kwok, R. C.-W., & Yue, W. T. (2023). Could Gamification Designs Enhance Online Learning Through Personalization? Lessons from a Field Experiment. *Information Systems Research*, *34*(1), 27–49. https://doi.org/10.1287/isre.2022.1123

Levitt, S. D., & List, J. A. (2007). What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *The Journal of Economic Perspectives*, *21*(2), 153–174. http://www.jstor.org/stable/30033722

Levitt, S. D., & List, J. A. (2008). Homo economicus Evolves. *Science*, *319*, 909–910. https://api.semanticscholar.org/CorpusID:155819315

Lieberman, H., Paternò, F., Klann, M., & Wulf, V. (2006, October). End-User Development: An Emerging Paradigm. In *End user development* (pp. 1–8). Springer Netherlands. https://doi.org/10.1007/1-4020-5386-x{_}1

Lim, J.-E., Lee, J., & Kim, D. (2021). The Effects of Feedback and Goal on the Quality of Crowdsourcing Tasks. *International Journal of Human–Computer Interaction*, *37*(13), 1207–1219. https://doi.org/10.1080/10447318.2021.1876355

Lindquist, J. D., & Kaufman-Scarborough, C. (2007). The Polychronic—Monochronic Tendency Model. *http://dx.doi.org/10.1177/0961463X07080270*, *16*(3), 253–285. https://doi.org/10.1177/0961463X07080270

Liu, Y. (2003). Developing a scale to measure the interactivity of websites. *Journal of Advertising Research*, *43*(2), 207–216. https://doi.org/10.1017/S0021849903030204

Luther, K., Pavel, A., Wu, W., Tolentino, J. L., Agrawala, M., Hartmann, B., & Dow, S. (2014). CrowdCrit: Crowdsourcing and Aggregating Visual Design Critique. *CSCW Companion '14: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 21–24. https://doi.org/10.1145/2556420.2556788

Luther, K., Tolentino, J. L., Wu, W., Pavel, A., Bailey, B. P., Agrawala, M., Hartmann, B., & Dow, S. P. (2015). Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 473–485. https://doi.org/10.1145/2675133.2675283

Ma, X., Li, Y., Forlizzi, J., & Dow, S. (2015). Exiting the Design Studio: Leveraging Online Participants for Early-Stage Design Feedback. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 676–685. https://doi.org/10.1145/2675133.2675174

Maalej, W., & Pagano, D. (2011). On the Socialness of Software. *Proceedings - IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, DASC 2011*, 864–871. https://doi.org/10.1109/DASC.2011.146

Mackay, W. E. (2004). The Interactive Thread: Exploring Methods for Multi-Disciplinary Design. *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '04)*, 103–112. https://doi.org/10.1145/1013115.1013131

Mao, J. Y., Vredenburg, K., Smith, P. W., & Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM*, *48*(3), 105–109. https://doi.org/10.1145/1047671.1047677

Mao, K., Capra, L., Harman, M., & Jia, Y. (2017). A survey of the use of crowdsourcing in software engineering. *Journal of Systems and Software*, *126*, 57–84. https://doi.org/10.1016/j.jss.2016.09.015

McFarlane, D. C., & Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, *17*(1), 1–61. https://doi.org/10.1207/S15327051HCI1701{_}1

McKeen, J. D., & Guimaraes, T. (1997). Successful Strategies for User Participation in Systems Development. *Journal of Management Information Systems*, *14*(2), 133–150. http://www.jstor.org/stable/40398269

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/https://doi.org/10.1016/S1364-6613(03)00028-7

Morgan, D. L. (1997). *Focus Groups as Qualitative Research*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984287

Morschheuser, B., Hamari, J., Koivisto, J., & Maedche, A. (2017). Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human Computer Studies*, *106*, 26–43. https://doi.org/10.1016/j.ijhcs.2017.04.005

Muñante, D., Siena, A., Kifetew, F. M., Susi, A., Stade, M., & Seyff, N. (2017). Gathering Requirements for Software Configuration from the Crowd. *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)*, 176–181. https://doi.org/10.1109/REW.2017.74

Murthi, B. P. S., & Sarkar, S. (2003). The Role of the Management Sciences in Research on Personalization. *Management Science*, *49*(10), 1344–1362. https://doi.org/10.1287/mnsc.49.10.1344.17313

Naudet, Y., & Lykourentzou, I. (2014). Personalisation in Crowd Systems. *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, 32–37. https://doi.org/10.1109/SMAP.2014.13

Nebeling, M., Speicher, M., & Norrie, M. C. (2013). CrowdStudy: General Toolkit for Crowdsourced Evaluation of Web Interfaces. *EICS '13: Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 255. https://doi.org/10.1145/2494603.2480303

Newman, G. E., & Cain, D. M. (2014). Tainted Altruism: When Doing Some Good Is Evaluated as Worse Than Doing No Good at All. *Psychological Science*, *25*(3), 648–655. https://doi.org/10.1177/0956797613504785

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *14769344*, *22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann Publishers Inc.

Oppenlaender, J., & Hosio, S. (2019). Towards Eliciting Feedback for Artworks on Public Displays. *C&C 19: Proceedings of the 2019 Conference on Creativity and Cognition*, 562–569. https://doi.org/10.1145/3325480.3326583

Oppenlaender, J., Kuosmanen, E., Lucero, A., & Hosio, S. (2021). Hardhats and Bungaloos: Comparing Crowdsourced Design Feedback with Peer Design Feedback in the Classroom. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445380

Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P., & Hosio, S. (2020). Creativity on Paid Crowdsourcing Platforms. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376677

Oppenlaender, J., Tiropanis, T., & Hosio, S. (2020). CrowdUI: Supporting Web Design with the Crowd. *Proceedings of the ACM on Human-Computer Interaction*, *4*(EICS), 1–28. https://doi.org/10.1145/3394978

Organisciak, P., Teevan, J., Dumais, S., Miller, R. C., & Kalai, A. T. (2015). Matching and Grokking: Approaches to Personalized Crowdsourcing. *Proceedings of the 24th International Conference on Artificial Intelligence*, 4296–4302.

Pagano, D. (2013). *PORTNEUF - A Framework for Continuous User Involvement* [Doctoral dissertation, Technical University Munich].

Pagano, D., & Bruegge, B. (2013). User Involvement in Software Evolution Practice: A Case Study. *2013 35th International Conference on Software Engineering*, 953–962. https://doi.org/10.1109/ICSE.2013.6606645

Paulino, D., Correia, A., Barroso, J., & Paredes, H. (2023). Cognitive Personalization for Online Microtask Labor Platforms: A Systematic Literature Review. *User Modeling and User-Adapted Interaction*, *2023*, 1–42. https://doi.org/10.1007/S11257-023-09383-W

Paulino, D., Correia, A., Guimarães, D., Barroso, J., & Paredes, H. (2022). Uncovering the Potential of Cognitive Personalization for UI Adaptation in Crowd Work. *2022*

IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 484–489. https://doi.org/10.1109/CSCWD54268.2022.9776164

Paulino, D., Guimaraes, D., Correia, A., Ribeiro, J., Barroso, J., & Paredes, H. (2023). A model for cognitive personalization of microtask design. *Sensors*, *23*, 3571. https://doi.org/10.3390/s23073571

Pedersen, J., Kocsis, D., Tripathi, A., Tarrell, A., Weerakoon, A., Tahmasbi, N., Xiong, J., Deng, W., Oh, O., & De Vreede, G. J. (2013). Conceptual Foundations of Crowdsourcing: A Review of IS Research. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 579–588. https://doi.org/10.1109/HICSS.2013.143

Piasentin, K. A., & Chapman, D. S. (2007). Perceived similarity and complementarity as predictors of subjective person-organization fit. *Journal of Occupational and Organizational Psychology*, *80*(2), 341–354. https://doi.org/10.1348/096317906X115453

Poposki, E. M., & Oswald, F. L. (2010). The Multitasking Preference Inventory: Toward an Improved Measure of Individual Differences in Polychronicity. *Human Performance*, *23*(3), 247–264. https://doi.org/10.1080/08959285.2010.487843

Pretel, I., Lopez-Novoa, U., Sanz-Yagüe, E., López-de-Ipiña, D., Cartelli, V., Di Modica, G., & Tomarchio, O. (2017). Citizenpedia: A human computation framework for the e-government domain. *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation*, 1–6. https://doi.org/10.1109/UIC-ATC.2017.8397542

Randall, T., Terwiesch, C., & Ulrich, K. T. (2007). Research Note—User Design of Customized Products. *Marketing Science*, *26*(2), 268–280. https://doi.org/https://doi.org/10.1287/mksc.1050.0116

Richman, A. L., Civian, J. T., Shannon, L. L., Jeffrey Hill, E., & Brennan, R. T. (2008). The Relationship of Perceived Flexibility, Supportive Work-Life Policies, and Use of Formal Flexible Arrangements and Occasional Flexibility to Employee Engagement and Expected Retention. *Community, Work and Family*, *11*(2), 183–197. https://doi.org/10.1080/13668800802050350

Rissler, R., Nadj, M., Adam, M., & Maedche, A. (2017). Towards an integrative Theoretical Framework of IT-Mediated Interruptions. *Proceedings of the 25th European*

Conference on Information Systems (ECIS 2017), 1950–1967. https://aisel.aisnet.org/ecis2017_rp/125

Robb, D. A., Padilla, S., Kalkreuter, B., & Chantler, M. J. (2015a). Crowd Sourced Feedback with Imagery Rather than Text: Would Designers Use It? *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, 1355–1364. https://doi.org/10.1145/2702123.2702470

Robb, D. A., Padilla, S., Kalkreuter, B., & Chantler, M. J. (2015b). Moodsource: Enabling Perceptual and Emotional Feedback from Crowds. *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, 21–24. https://doi.org/10.1145/2685553.2702676

Robb, D. A., Padilla, S., Methven, T. S., Kalkreuter, B., & Chantler, M. J. (2017). Image-Based Emotion Feedback: How does the Crowd Feel? And Why? *DIS 2017 - Proceedings of the 2017 ACM Conference on Designing Interactive Systems*, 451–463. https://doi.org/10.1145/3064663.3064665

Roetzel, P. G. (2019). Information Overload in the Information Age: A Review of the Literature from Business Administration, Business Psychology, and Related Disciplines with a Bibliometric Approach and Framework Development. *Business Research*, *12*(2), 479–522. https://doi.org/10.1007/s40685-018-0069-z

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2021). An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *Proceedings of the International AAAI Conference on Web and Social Media*, 321–328. https://doi.org/10.1609/icwsm.v5i1.14105

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/https://doi.org/10.1006/ceps.1999.1020

Saito, S., Nakano, T., Chiang, C. W., Kobayashi, T., Savage, S., & Bigham, J. P. (2019). TurkScanner: Predicting the Hourly Wage of Microtasks. *Proceedings of the World Wide Web Conference (WWW 2019)*, 3187–3193. https://doi.org/10.1145/3308558.3313716

Sarı, A., Tosun, A., & Alptekin, G. I. (2019). No Title. *Journal of Systems and Software*, *153*, 200–219. https://doi.org/10.1016/j.jss.2019.04.027

Sarstedt, M., & Mooi, E. (2014). *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics* (2nd ed. 20). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12541-6

Schader, M., Geiger, D., Rosemann, M., & Fielt, E. (2012). *Crowdsourcing Information Systems - Definition, Typology, and Design* (Vol. 4). https://aisel.aisnet.org/icis2012/proceedings/ResearchInProgress/53

Schneider, H., Frison, K., Wagner, J., & Butz, A. (2016). CrowdUX: A Case for Using Widespread and Lightweight Tools in the Quest for UX. *DIS 2016 - Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 415–425. https://doi.org/10.1145/2901790.2901814

Scholtz, J. (2001). Adaptation of Traditional Usability Testing Methods for Remote Testing. *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, 5030. https://doi.org/10.1109/HICSS.2001.926546

Schulze, T., Schader, M., & Krug, S. (2012). Workers' Task Choice in Crowdsourcing and Human Computation Markets. *International Conference on Information Systems (ICIS 2012)*, *4*, 3551–3561.

Scriven, M. (1991). Beyond Formative and Summative Evaluation. *Evaluation and education: At quarter century*, *10*(Part II), 19–64. https://doi.org/10.1177/016146819109200603

Sekiguchi, T. (2004). Person-Organization Fit and Person-Job Fit in Employee Selection: A Review of the Literature. *Osaka Keidai Ronshu*, *54*, 179–196. https://api.semanticscholar.org/CorpusID:114928430

Seyff, N., Graf, F., & Maiden, N. (2010). Using Mobile RE Tools to Give End-Users their Own Voice. *Proceedings of the 2010 18th IEEE International Requirements Engineering Conference, RE2010*, 37–46. https://doi.org/10.1109/RE.2010.15

Seyff, N., Ollmann, G., & Bortenschlager, M. (2014). AppEcho: A User-Driven, In Situ Feedback Approach for Mobile Platforms and Applications. *Proceedings of the 1st International Conference on Mobile Software Engineering and Systems*, 99–108. https://doi.org/10.1145/2593902.2593927

Sherief, N., Jiang, N., Hosseini, M., Phalp, K., & Ali, R. (2014). Crowdsourcing Software Evaluation. *Proceedings of the 18th International Conference on Evaluation and*

Assessment in Software Engineering, 1–4. https://doi.org/10.1145/2601248.2601300

Snijders, R., Dalpiaz, F., Brinkkemper, S., Hosseini, M., Ali, R., & Ozum, A. (2015). REfine: A Gamified Platform for Participatory Requirements Engineering. *2015 IEEE 1st International Workshop on Crowd-Based Requirements Engineering (CrowdRE)*, 1–6. https://doi.org/10.1109/CrowdRE.2015.7367581

Stade, M., Oriol, M., Cabrera, O., Fotrousi, F., Schaniel, R., Seyff, N., & Schmidt, O. (2017). Providing a User Forum is not Enough: First Experiences of a Software Company with CrowdRE. *Proceedings - 2017 IEEE 25th International Requirements Engineering Conference Workshops, REW 2017*, 164–169. https://doi.org/10.1109/REW.2017.21

Staiger, A.-M., Schmidt, J., & von der Oelsnitz, D. (2022). How Well did I Do? The Effect of Feedback on Affective Commitment in the Context of Microwork. *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2022.636

Stangl, A., Morris, M. R., & Gurari, D. (2020). "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376404

Stangl, A., Verma, N., Fleischmann, K. R., Morris, M. R., & Gurari, D. (2021). Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who Are Blind or Have Low Vision. https://doi.org/10.1145/3441852.3471233

Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2017). Toward a Theory of Interactive Media Effects (TIME). In *The handbook of the psychology of communication technology* (pp. 47–86). https://doi.org/10.1002/9781118426456.ch3

Sundar, S. S., & Marathe, S. S. (2010). Personalization versus Customization: the Importance of Agency, Privacy, and Power Usage. *Human Communication Research*, *36*(3), 298–322. https://doi.org/10.1111/j.1468-2958.2010.01377.x

Sykes, T. A. (2020). Enterprise System Implementation and Employee Job Outcomes: Understanding the Role of Formal and Informal Support Structures Using the Job

Strain Model. *MIS Quaterly*, *44*(4), 2055–2086. https://doi.org/10.25300/MISQ/2020/11672

Tam, K. Y., & Ho, S. Y. (2006). Understanding the Impact of Web Personalization on User Information Processing and Decision Outcomes. *MIS Quarterly*, *30*(4), 865–890. https://doi.org/10.2307/25148757

Tams, S., Thatcher, J., & Grover, V. (2018). Concentration, Competence, Confidence, and Capture: An Experimental Study of Age, Interruption-based Technostress, and Task Performance. *Journal of the Association for Information Systems*, *19*, 857–908. https://doi.org/10.17705/1jais.00511

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition* (4. ed.). Elsevier Science & Technology Books. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=320843

Thirumalai, S., & Sinha, K. K. (2013). To Personalize or Not to Personalize Online Purchase Interactions: Implications of Self-Selection by Retailers. *Information Systems Research*, *24*(3), 683–708. http://www.jstor.org/stable/42004288

Tizard, J., Rietz, T., Liu, X., & Blincoe, K. (2022). Voice of the Users: An Extended Study of Software Feedback Engagement. *Requirements Engineering*, *27*(3), 293–315. https://doi.org/10.1007/s00766-021-00357-1

Tonin, M., & Vlassopoulos, M. (2015). Corporate Philanthropy and Productivity: Evidence from an Online Real Effort Experiment. *Management Science*, *61*(8), 1795–1811. http://www.jstor.org/stable/24551507

Tremblay, M. C., Hevner, A. R. ; & Berndt, D. J. (2010). Focus Groups for Artifact Refinement and Evaluation in Design Research. *Communications of the Association for Information Systems*, *26*(1), 27. https://doi.org/10.17705/1CAIS.02627

van Griethuijsen, R. A., van Eijck, M. W., Haste, H., den Brok, P. J., Skinner, N. C., Mansour, N., Gencer, A. S., & BouJaoude, S. (2015). Global Patterns in Students' Views of Science and Interest in Science. *Research in Science Education*, *45*(4), 581–603. https://doi.org/10.1007/s11165-014-9438-6

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. https://doi.org/10.1057/ejis.2014.36

Venkatesh, V., Windeler, J. B., Bartol, K. M., & Williamson, I. O. (2017). Person-Organization and Person-Job Fit Perceptions of New It Employees: Work Outcomes and Gender

Differences. *MIS Quarterly*, *41*(2), 525–558. https://doi.org/10.25300/MISQ/2017/41.2.09

Vredenburg, K., Mao, J. Y., Smith, P. W., & Carey, T. (2002). A Survey of User-Centered Design Practice. *Proceedings of the 2002 CHI Conference on Human Factors in Computing Systems*, 471–478. https://doi.org/10.1145/503376.503460

Wang, J., Yang, Y., Wang, S., Chen, C., Wang, D., & Wang, Q. (2022). Context-Aware Personalized Crowdtesting Task Recommendation. *IEEE Transactions on Software Engineering*, *48*(8), 3131–3144. https://doi.org/10.1109/TSE.2021.3081171

Wang, J., Ipeirotis, P. G., & Provost, F. (2017). Cost-Effective Quality Assurance in Crowd Labeling. *Information Systems Research*, *28*(1), 137–158. https://doi.org/10.1287/isre.2016.0661

Warr, P., & Inceoglu, I. (2012). Job engagement, job satisfaction, and contrasting associations with person–job fit. *Journal of occupational health psychology*, *17*(2), 129.

Wauck, H., Yen, Y.-C. C., Fu, W.-T. T., Gerber, E., Dow, S. P., & Bailey, B. P. (2017). From in the Class or in the Wild? Peers provide better design feedback than external crowds. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 5580–5591. https://doi.org/10.1145/3025453.3025477

Webster, J., & Watson, R. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MISQ*, *6*(2), xiii–xxiii. http://www.jstor.org/stable/4132319

Webster, J., & Ho, H. (1997). Audience Engagement in Multimedia Presentations. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *28*(2), 63–77. https://doi.org/10.1145/264701.264706

Wecker, A. J., Schor, U., Raziel-Kretzmer, V., Elovits, D., Lavee, M., Kuflik, T., & Stoekl Ben Ezra, D. (2020). Opportunities for Personalization for Crowdsourcing in Handwritten Text Recognition. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 373–375. https://doi.org/10.1145/3386392.3402436

Wecker, A. J., Schor, U., Elovits, D., Stoekl Ben Ezra, D., Kuflik, T., Lavee, M., Raziel-Kretzmer, V., Ohali, A., & Signoret, L. (2019). Tikkoun Sofrim: A WebApp for Personalization and Adaptation of Crowdsourcing Transcriptions. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP'19 Adjunct)*. https://doi.org/10.1145/3314183.3324972

Weinmann, M., Hibbeln, M., & Robra-Bissantz, S. (2011). Customer-Oriented Configuration Systems: One Type Fits All? *19th European Conference on Information Systems, ECIS 2011*. https://aisel.aisnet.org/ecis2011/132

Whiting, M. E., Hugh, G., & Bernstein, M. S. (2019). Fair Work: Crowd Work Minimum Wage with One Line of Code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 197–206. https://ojs.aaai.org/index.php/HCOMP/article/view/5283

Williams, A. C., Mark, G., Milland, K., Lank, E., & Law, E. (2019). The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–28. https://doi.org/10.1145/3359126

Williams, C., De Greef, L., Harris, E., Findlater, L., Pavel, A., & Bennett, C. (2022). Toward Supporting Quality Alt Text in Computing Publications. *Proceedings of the 19th International Web for All Conference, W4A 2022*, 12. https://doi.org/10.1145/3493612.3520449

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. *Proceedings of the 2011 CHI Conference Conference on Human Factors in Computing Systems*, 143–146. https://doi.org/10.1145/1978942

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013). Using Grounded Theory as a Method for Rigorously Reviewing literature. https://doi.org/10.1057/ejis.2011.51

Wu, M.-H., & Quinn, A. J. (2017). Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. *AAAI Conference on Human Computation \& Crowdsourcing*. https://api.semanticscholar.org/CorpusID:34384062

Wu, Y. W., & Bailey, B. P. (2016). Novices Who Focused or Experts Who Didn't? In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4086–4097). Association for Computing Machinery (ACM). https://doi.org/10.1145/2858036.2858330

Wu, Y. W., & Bailey, B. P. (2021). Better Feedback from Nicer People: Narrative Empathy and Ingroup Framing Improve Feedback Exchange. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–20. https://doi.org/10.1145/3432935

Xu, A., & Bailey, B. P. (2011). A Crowdsourcing Model for Receiving Design Critique. *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*, 1183–1188. https://doi.org/10.1145/1979742.1979745

Xu, A., & Bailey, B. P. (2014). A System for Receiving Crowd Feedback on Visual Designs Abstract. *CSCW Companion '14: Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 37–40. https://doi.org/10.1145/2556420.2556791

Xu, A., Huang, S.-W. W., Bailey, B. P., & Duque-Estrada, A. (2014). Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, *24*(74), 1433–1444. https://doi.org/10.1145/2531602.2531604

Xu, A., Rao, H., Dow, S. P., & Bailey, B. P. (2015). A Classroom Study of Using Crowd Feedback in the Iterative Design Process. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1637–1648. https://doi.org/10.1145/2675133.2675140

Yen, Y. C. G., Dow, S. P., Gerber, E., & Bailey, B. P. (2016). Social network, Web Forum, or Task Market? Comparing Different Crowd Genres for Design Feedback Exchange. *DIS 2016 - Proceedings of the 2016 ACM Conference on Designing Interactive Systems: Fuse*, 773–784. https://doi.org/10.1145/2901790.2901820

Yen, Y. C. G., Dow, S. P., Gerber, E., & Bailey, B. P. (2017). Listen to Others, Listen to Yourself. *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 158–170. https://doi.org/10.1145/3059454.3059468

Yen, Y. C. G., Kim, J. O., & Bailey, B. P. (2020). Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, *20*, 1–13. https://doi.org/10.1145/3313831.3376380

Yu, Z., Xu, Z., Black, A., & Rudnicky, A. I. (2016). Chatbot Evaluation and Database Expansion via Crowdsourcing. *Proceedings of the Chatbot Workshop of LREC*, 1–5. https://api.semanticscholar.org/CorpusID:12117509

Yuan, A., Luther, K., Krause, M., Vennix, S. I., Dow, S. P., & Hartmann, B. B. (2016). Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. *Proceedings of the ACM Conference on Computer*

Supported Cooperative Work, CSCW, *27*, 1005–1017. https://doi.org/10.1145/2818048.2819953

Zhang, T., Agarwal, R., & Lucas, H. C. (2011). The Value of It-Enabled Retailer Learning: Personalized Product Recommendations and Customer Store Loyalty in Electronic Markets. *MIS Quarterly*, *35*(4), 859–881. https://doi.org/10.2307/41409964

Zuchowski, O., Posegga, O., Schlagwein, D., & Fischbach, K. (2016). Internal crowdsourcing: Conceptual framework, structured review, and research agenda. *Journal of Information Technology*, *31*(2), 166–184. https://doi.org/10.1057/jit.2016.14

Zwicky, F., & Wilson, A. G. (1967). *New Methods of Thought and Procedure: Contributions to the Symposium on Methodologies*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-87617-2

Appendix

A. Appendix for Study I

Stream	Author	Input						Crowd Configuration														Design Characteristics										Effects		
		IS Lifecycle Stage		Type of Feedback		Feedback Scope		Crowd Type				Incentive								Feedback Collection Mechanisms						Interactivity Cues								
		Development	Operations	Qualitative	Quantitative	Non-functional Attributes	Content	Functional Attributes	Anonymous	Proxy Users	Students	Convenience	Money	Improvement/ Involvement	Interest/Social Compensation	Course Credit	Gamification	Questionnaire	Free Text Field	Categories	Selection	Direct Manipulation	Collaboration	Marker	Context	Recording	Outcome Effects	Process Effects	Intermediate Effects
1 – Anonymous Crowd-Feedback	Greenberg et al. 2015	X		X		X	X		X				X														X	X				
	Kang et al. 2018	X		X		X	X		X				X							X							X			X			X
	Krause et al. 2017	X		X		X	X		X				X						X								X						
	Lekschas et al. 2021	X		X		X	X		X				X		X					X								X					
	Luther et al. 2014; Luther et al. 2015; Yuan et al. 2016	X		X		X			X				X							X								X					
	Wauck et al. 2017	X		X		X	X		X		X	X	X		X	X			X								X	X					
	Wu and Bailey 2016	X		X		X	X		X				X		X				X						X		X						
	Wu and Bailey 2021	X	X	X		X	X		X				X						X								X						
	Xu et al. 2014; Xu et al. 2015; Xu and Bailey 2014	X		X	X	X	X		X				X							X								X					
	Yen et al. 2016	X		X		X	X		X			X	X		X				X											X			
	Yen et al. 2017	X		X		X			X				X							X								X					
2 – Real User Crowd-Feedback	Easterday et al. 2017	X		X	X		X	X		X										X								X			X		
	Haukipuro et al. 2016	X		X	X			X		X				X					X				X						X		X		
	Muñante et al. 2017	X		X	X			X		X									X						X				X				
	Nebeling et al. 2013	X	X	X	X	X	X	X	X	X			X	X					X				X						X	X			
	Oppenlaender et al. 2020		X	X	X	X			X					X								X	X					X	X				
	Pretel et al. 2017		X	X				X		X				X				X		X			X										
	Schneider et al. 2016	X		X	X	X	X	X		X			X								X			X				X	X				
	Seyff et al. 2014		X	X						X				X					X						X		X	X					
	Snijders et al. 2015	X	X	X	X			X		X				X		X			X				X			X			X				X
	Stade et al. 2017		X	X	X	X	X	X		X				X						X				X	X				X				
3 – Hybrid Crowd-Feedback	Ayalon and Toch 2018		X		X	X			X				X					X								X			X				
	Ayalon and Toch 2019	X			X	X			X				X					X								X				X			
	Choi et al. 2020	X		X			X	X	X				X								X	X							X	X			
	La Cruz et al. 2015	X	X		X			X	X				X								X								X	X			
	Dow et al. 2013	X		X		X		X		X			X					X							X					X			
	Hosseini et al. 2016		X		X	X				X				X							X									X		X	
	Jansson and Bremdal 2018	X			X	X															X									X		X	
	Ma et al. 2015	X		X		X		X	X				X						X							X				X			
	Oppenlaender and Hosio 2019	X		X	X	X				X				X							X				X					X			X
	Oppenlaender et al. 2021	X		X	X	X	X	X	X			X				X		X												X			
	Robb et al. 2015a, 2015b; Robb et al. 2017	X		X		X	X		X		X			X			X				X									X		X	
	Xu and Bailey 2011	X		X		X			X									X						X									
	Yu et al. 2016	X		X	X		X	X	X				X								X									X			X
Sum		**28**	**10**	**29**	**17**	**25**	**17**	**15**	**21**	**13**	**3**	**2**	**22**	**8**	**5**	**3**	**2**	**10**	**8**	**8**	**7**	**2**	**7**	**7**	**4**	**4**	**22**	**16**	**7**				
Percentage		**78%**	**26%**	**81%**	**47%**	**68%**	**47%**	**42%**	**58%**	**36%**	**8%**	**6%**	**61%**	**22%**	**14%**	**8%**	**6%**	**28%**	**22%**	**22%**	**19%**	**6%**	**19%**	**19%**	**11%**	**11%**	**61%**	**44%**	**19%**				

Table A.1.: Concept Matrix for Identified Articles Targeting Crowd-Feedback Systems

B. Appendix for Study II

Interview Guides

Interview Guide Study 1

1. Einstieg
    a) Wie haben Sie die Arbeit mit dem Feedback-Tool empfunden?
    b) Welche Vor- und Nachteile sehen Sie im Gegensatz zu einem persönlichen Interview?
2. Feedbackmenge
    a) Haben Sie noch Feedback zum Prototyp, das sie nicht im Experiment angegeben haben?
        i. Wenn ja, warum haben Sie das Feedback nicht angegeben?
    b) Fällt Ihnen eine Funktion ein, die Ihnen hätte helfen können, mehr Feedback anzugeben?
    c) Wie unterschiedet sich Ihr Feedback von dem Feedback, das Sie in einem persönlichen Interview gegeben hätten?
3. Feedbackstruktur
    a) Fanden Sie die Vorstrukturierung nützlich oder eher hinderlich?
    b) Wie sind Sie mit den Erklärungen zu den Kategorien zurechtgekommen?
    c) Was könnte man an der Feedbackstruktur noch verbessern?
4. Allgemeines
    a) Wie sollte man aus Ihrer Sicht Feedback-Tools für das bezahlte Crowdsourcen von Feedback designen?
    b) Wussten Sie zu jedem Zeitpunkt, was Sie als nächstes tun sollten?
    c) Haben Sie noch weitere Ideen, wie man das Feedback-Tool verbessern könnte?
    d) Wie würden Sie conversational Feedback empfinden? Wie sollte natürliche Sprache integriert werden?
    e) Was halten Sie von einer Spracheingabe für Feedback?

Interview Guide Study 2

1. Overall Experience
    a) How was your overall experience with Feeasy?
        i. What did you like about Feeasy?
        ii. What did you not like about Feeasy?
    b) How did feature xyz enabled you to provide better feedback?
    c) Why did you use or not use feature xyz?
2. Perceived Effects on Feedback Quality
    a) Did you feel able to address specific elements of the user interface?
    b) How did you manage to make sure that it is clear which element of the user interface your feedback comment is addressing?
    c) How did you identify design issues?
    d) How did you decide which feedback is relevant? Did you feel confident in deciding which topics are relevant?
    e) How did you establish the desired level of objectivity? Did you feel able to identify design issues?
    f) Do you think your feedback is rather objective or subjective?
3. Outlook
    a) What other features would have helped you to provide better feedback?
    b) In which situations do you think the feature xyz is helpful to provide better feedback?

Constructs and Items

Construct	Items	Reference
Perceived Interactivity	I felt that I had a lot of control over my visiting experiences at Feeasy While I was on Feeasy, I could choose freely what I wanted to see While surfing Feeasy, I had absolutely no control over what I can do on the site While surfing Feeasy, my actions decided the kind of experiences I got Feeasy makes me feel it wants to listen to its users Feeasy does not at all encourage users to talk back Feeasy gives users the opportunity to talk back	Adapted from Liu (2003)
Perceived User Engagement	Feeasy kept me totally absorbed in the browsing Feeasy held my attention Feeasy excited my curiosity Feeasy aroused my imagination Feeasy was fun Feeasy was intrinsically interesting Feeasy was engaging	Adapted from Webster and Ho (1997)
Perceived Ease of Use	Learning to operate Feeasy was easy for me I found it easy to get Feeasy to do what I want it to do My interaction with Feeasy was clear and understandable I found Feeasy to be flexible to interact with It was easy for me to become skillful at using Feeasy I found Feeasy easy to use	Adapted from Davis (1989)

Table B.2.: Items Study 2

C. Appendix for Study IV

Interview Guide

1. Overall Experience

    a) How was your experience using the CrowdSurfer?

    b) What are the differences between working with the CrowdSurfer and doing traditional crowdworking tasks?

    c) How did the CrowdSurfer impact how comfortable you felt doing crowdwork?

    d) Which functionalities did you use?

    e) What are the key advantages of the CrowSurfer?

    f) What are the key disadvantages of the CrowdSurfer?

    g) How could the CrowdSurfer be improved?

2. Feedback Process

a) How did you experience the feedback tasks?

b) When did you provide feedback?

c) What motivated you to provide feedback?

d) Please reimagine the feedback provision process with the CrowdSurfer. How did you decide if you want to provide feedback?

e) Are there specific situations in which you did or did not provide feedback?

f) Need: How much do you think your feedback was wanted?

g) Ability: How much do you think you had the ability to provide meaningful feedback?

h) Effective Action: How much did you think you could make an impact by providing feedback?

3. Invisible Work

a) How do you think the CrowdSurfer impacts the amount of invisible work you have when doing tasks?

b) How do you think the CrowdSurfer impacts your work life balance?

c) How do you think the CrowdSurfer impacts your flexibility of doing crowdworking tasks?

4. Further Ideas

a) Can you remember other tasks for which such a plugin could be helpful?

Constructs and Items

Construct	Items	Reference
Perceived Task Completion Time	I felt that I had a lot of control over my visiting experiences at Feeasy	Self-developed
Fairness of Payment	The hourly payment for this task is appropriate The reward for this task is reasonable Compared to other tasks on the platform, the payment for this task is okay	Adapted from Schulze et al. (2012)
Work Flexibility	I can decide, when I want to work I can define my working pace myself I can contact the requester at any time There are different possibilities of contacting the requester I have a say regarding the focus of the tasks I can prioritise tasks in my working I can work on tasks of special interest	Adapted from Richman et al. (2008) and Kokoç (2019)

Table C.3.: Items Study 4

D. Appendix for Study V

	Selfish Instantiation	Altruistic Instantiation
Bonus task instructions	Thank you for also working on the bonus task to make Wikipedia more accessible! Of course, this task allows you to earn some extra money. At the same time, your work will make pictures accessible. Every picture will add one more data point for the project. We ask you to do at least the mandatory main tasks. However, we expect you to do your best and contribute picture annotations to the project. The more pictures you manage to annotate, the better for the project and the more bonus money you will earn. This is your chance to earn extra money by doing an exceptional job! Thank you for participating in the project and working on the bonus task to annotate pictures!	Thank you for also being part of our mission to make Wikipedia more accessible! So, why should you engage in our bonus task? You might think, "to earn some extra money", but in reality, you are doing something special. Your efforts will help to make pictures accessible for visually impaired readers. Every picture count and brings us closer to the goal of tearing down the barriers for visually impaired readers. We ask you to do at least the mandatory main tasks. However, we also invite you to do your very best and contribute as much alt-tags as you can. The more pictures you annotate, the better! Thank you for "being the eyes" of others and helping out others in need!
Info on submitted bonus tasks	Bonus earned via alt-tags	Alt-tags contributed for accessibility
Feedback message	You are doing an **exceptional job**! By **doing your best** and annotating more pictures, you can earn **bonus money** and contribute more to our project.	Thank you for contributing to make Wikipedia **more accessible**! Every picture counts and allows more **visually impaired** people to experience Wikipedia **without any barriers**.

Table D.4.: Overview of differences between the selfish and altruistic instantiation

Construct	Items	Reference
Polychronicity (Multitaskign Preference Inventory (MPI))	I prefer to work on several projects in a day, rather than completing one project and then switching to another.	Poposki and Oswald (2010)
	I would like to work in a job where I was constantly shifting from one task to another, like a receptionist or an air traffic controller.	
	I lose interest in what I am doing if I have to focus on the same task for long periods of time, without thinking about or doing something else.	
	When doing a number of assignments, I like to switch back and forth between them rather than do one at a time.	
	I like to finish one task completely before focusing on anything else. (R)	
	It makes me uncomfortable when I am not able to finish one task completely before focusing on another task. (R)	
	I am much more engaged in what I am doing if I am able to switch between several different tasks.	
	I do not like having to shift my attention between multiple tasks. (R)	
	I would rather switch back and forth between several projects than concentrate my efforts on just one.	
	I would prefer to work in an environment where I can finish one task before starting the next. (R)	
	I don't like when I have to stop in the middle of a task to work on something else. (R)	
	When I have a task to complete, I like to break it up by switching to other tasks intermittently.	
	I have a "one-track" mind. (R)	
	I prefer not to be interrupted when working on a task. (R)	
Altruism	Imagine the following situation: Today you unexpectedly received 1,600 U.S. dollars. How much of this amount would you donate to a good cause? ($0 - $1,600)	Falk, Becker, Dohmen, Enke, et al. (2018) and Falk, Becker, Dohmen, Huffman, and Sunde (2023)
	How willing are you to give to good causes without expecting anything in return? (0 - 10)	
Person-Job Fit	I fitted right in to the job.	Venkatesh et al. (2017)
	Taking everything into account, the job was a complete fit for me.	
	The job provided a total fit for me.	
Job Satisfaction	Overall, I was satisfied with my job.	Sykes (2020)
	I would have preferred another, more ideal job. (R)	
	I was satisfied with the important aspects of my job.	
Job Performance	I managed to plan my work so that it was done on time.	Koopmans et al. (2014)
	My planning was optimal.	
	I kept in mind the results that I had to achieve in my work.	
	I was able to separate main issues from side issues at work.	
	I was able to perform my work well with minimal time and effort.	
	I took on extra responsibilities.	
	I started new tasks myself, when my old ones were finished.	
	I took on challenging work tasks, when available.	
	I worked at keeping my job knowledge up-to-date.	
	I worked at keeping my job skills up-to-date.	
	I came up with creative solutions to new problems.	
	I kept looking for new challenges in my job.	

Table D.5.: Items of Study 5 in the Pre-Screening and Post-Task Questionnaire

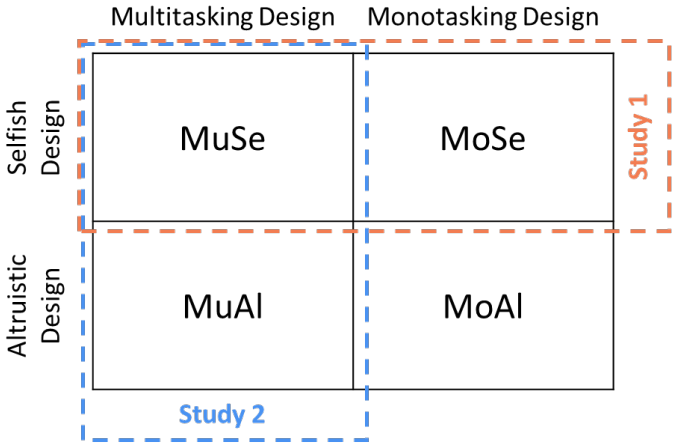


Figure D.1.: Combinations of design options for the study artifacts of study 1 and 2

Variable	NoFit (n = 32)	Fit (n = 44)	Analysis results
Age	35.28 (11.30)	36.14 (13.79)	Not sign., p = 0.84
Female	0.406 (0.499)	0.591 (0.497)	Not sign., p = 0.17
US	0.344 (0.483)	0.273 (0.451)	Not sign., p = 0.68
UK	0.375 (0.492)	0.364 (0.487)	Not sign., p = 1
ZA	0.281 (0.457)	0.364 (0.487)	Not sign., p = 0.61

Note: p-values are based on two-sided Mann-Whitney U tests. Two participants were not included because they answered ”diverse” when asked about their gender.

Table D.6.: Study 1: Means of key demographics over NoFit vs Fit

Variable	NoFit (n = 34)	Fit (n = 44)	Analysis results
Polychronicity	42.85 (10.70)	37.61 (13.54)	Sign., p = 0.06
Altruism	0.138 (0.799)	-0.0701 (0.998)	Not sign., p = 0.37

Note: p-values are based on two-sided Mann-Whitney U tests.

Table D.7.: Study 1: Means of preferences over NoFit vs Fit

Variable	NoFit (n = 34)	Fit (n = 44)	Analysis results
Job Satisfaction	5.059 (1.017)	4.682 (1.300)	Not sign., p = 0.20
Person-Job fit	4.775 (1.517)	4.326 (1.801)	Not sign., p = 0.37
Fairness of Payment	4.716 (1.525)	4.038 (1.779)	Sign., p = 0.09

Note: p-values are based on two-sided Mann-Whitney U tests.

Table D.8.: Study 1: Means of job satisfaction, person-job fit and fairness of payment for the bonus task over NoFit vs Fit

Dep. Var.: Quantity	(1)	(2)	(3)	(4)
Fit	-4.606	-4.289	-2.594	-8.817
	(8.019)	(8.074)	(7.914)	(10.747)
Multitasking		9.867	9.549	0.814
		(8.070)	(7.928)	(13.692)
Female			2.008	2.135
			(7.336)	(7.339)
Age			-0.538	-0.530
			(0.353)	(0.361)
UK			-1.278	-2.028
			(9.905)	(10.251)
ZA			-8.867	-7.409
			(10.554)	(10.855)
Fit x Multitasking				14.884
				(18.223)
Constant	33.765***	29.411***	49.789***	52.871***
	(6.442)	(7.625)	(17.400)	(18.572)
R^2	0.005	0.025	0.056	0.067
Observations	78	78	76	76

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.9.: Study 1: OLS regressions with *Quantity* as the dependent variable and all coefficients

Dep. Var.: $Quantity_{adj.}$	(1)	(2)	(3)	(4)
Fit	-0.450	-0.280	0.668	-1.045
	(4.251)	(4.246)	(4.231)	(5.468)
Multitasking		5.305	4.726	2.322
		(4.350)	(4.203)	(7.305)
Female			1.993	2.028
			(3.957)	(3.968)
Age			-0.219	-0.217
			(0.206)	(0.209)
UK			1.531	1.325
			(5.684)	(5.898)
ZA			-4.897	-4.496
			(5.070)	(5.242)
Fit x Multitasking				4.097
				(9.931)
Constant	16.100***	13.759***	21.025**	21.873**
	(3.330)	(3.627)	(9.169)	(9.698)
R^2	0.000	0.021	0.045	0.048
Observations	78	78	76	76

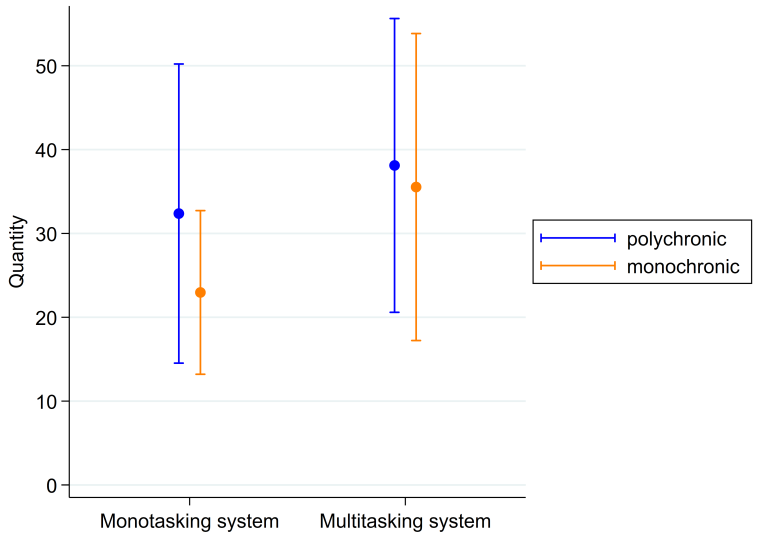
Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.10.: Study 1: OLS regressions with $Quantity_{adjusted}$ as dependent variable and all coefficients

Dep. Var.: Relevance	(1)	(2)	(3)	(4)
Fit	0.053	0.048	0.054	0.051
	(0.038)	(0.037)	(0.040)	(0.045)
Multitasking		-0.057	-0.055	-0.059
		(0.038)	(0.038)	(0.067)
Female			-0.053	-0.052
			(0.040)	(0.040)
Age			0.001	0.001
			(0.002)	(0.002)
UK			0.076	0.075
			(0.054)	(0.056)
ZA			0.038	0.039
			(0.042)	(0.042)
Fit x Multitasking				0.007
				(0.083)
Constant	0.624***	0.653***	0.606***	0.607***
	(0.030)	(0.031)	(0.065)	(0.063)
R^2	0.030	0.066	0.148	0.149
Observations	67	67	65	65

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.11.: Study 1: OLS regressions with Relevance (quality) as dependent variable and all coefficients



Note: Dots indicate means and whiskers indicate 95% confidence intervals.

Figure D.2.: Study 1: *Quantity* over Monotasking vs. Multitasking instantiation and polychronic vs. monochronic

Variable	NoFit (n = 33)	Fit (n = 35)	Analysis results
Age	32.48 (10.18)	33.89 (11.21)	Not sign., p = 0.73
Female	0.515 (0.508)	0.514 (0.507)	Not sign., p = 1
US	0.182 (0.392)	0.343 (0.482)	Not sign., p = 0.22
UK	0.455 (0.506)	0.257 (0.443)	Not sign., p = 0.15
ZA	0.364 (0.489)	0.400 (0.497)	Not sign., p = 0.95

Note: p-values are based on two-sided Mann-Whitney U tests.

Table D.12.: Study 2: Means of key demographics over *NoFit* vs *Fit*

Variable	NoFit (n = 33)	Fit (n = 35)	Analysis results
Polychronicity	34.36 (11.34)	39.49 (11.95)	Sign., p = 0.09
Altruism	0.141 (0.910)	0.0250 (0.916)	Not sign., p = 0.52

Note: p-values are based on two-sided Mann-Whitney U tests.

Table D.13.: Study 2: Means of preferences over *NoFit* vs *Fit*

Variable	NoFit (n = 33)	Fit (n = 35)	Analysis results
Job Satisfaction	4.727 (1.473)	4.686 (1.391)	Not sign., p = 0.85
Person-Job fit	4.535 (1.873)	4.324 (1.821)	Not sign., p = 0.56
Fairness of Payment	3.990 (1.676)	4.648 (1.540)	Sign., p = 0.09

Note: p-values are based on two-sided Mann-Whitney U tests.

Table D.14.: Study 2: Means of job satisfaction, person-job fit and fairness of payment for the bonus task over *NoFit* vs *Fit*

Dep. Var.: Quantity	(1)	(2)	(3)	(4)
Fit	-11.223*	-10.609	-13.129*	-28.849***
	(6.605)	(6.688)	(6.876)	(9.293)
Selfish		6.680	3.045	-11.304
		(6.300)	(5.841)	(10.364)
Female			-20.096***	-20.626***
			(7.409)	(7.059)
Age			0.382	0.545*
			(0.321)	(0.315)
UK			-6.121	-7.356
			(8.863)	(9.078)
ZA			11.663	17.380
			(9.741)	(10.414)
Fit x Selfish				26.686*
				(13.769)
Constant	30.394***	26.345***	25.017*	27.191*
	(5.225)	(5.751)	(14.779)	(14.080)
R^2	0.042	0.057	0.206	0.253
Observations	68	68	68	68

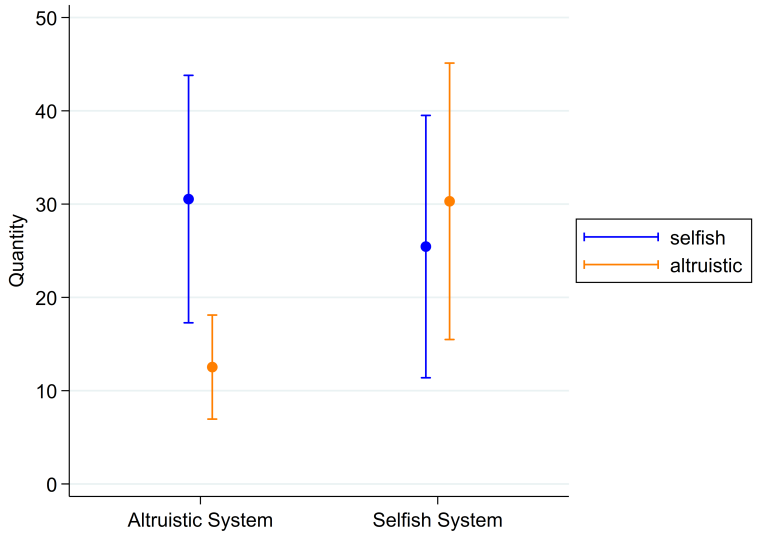
Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.15.: Study 2: OLS regressions with *Quantity* as dependent variable and all coefficients

Dep. Var.: $Quantity_{adj.}$	(1)	(2)	(3)	(4)
Fit	-6.029	-5.674	-6.925*	-18.278***
	(3.707)	(3.817)	(3.870)	(5.217)
Selfish		3.869	1.754	-8.608
		(3.653)	(3.443)	(5.850)
Female			-11.313***	-11.695***
			(4.234)	(3.966)
Age			0.225	0.342*
			(0.203)	(0.193)
UK			-2.703	-3.595
			(5.158)	(5.170)
ZA			5.460	9.589*
			(5.466)	(5.541)
Fit x Selfish				19.272**
				(7.408)
Constant	16.403***	14.058***	13.100	14.670
	(2.866)	(3.588)	(9.479)	(8.861)
R^2	0.039	0.055	0.191	0.269
Observations	68	68	68	68

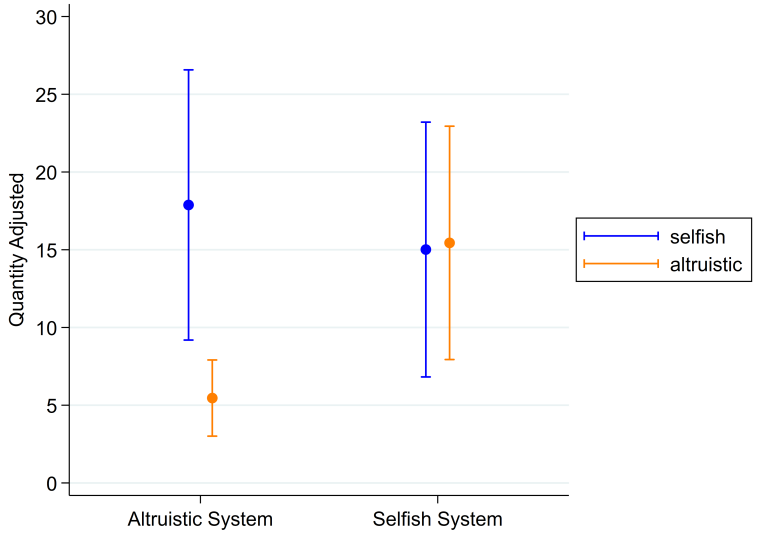
Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.16.: Study 2: OLS regressions with $Quantity_{adjusted}$ as dependent variable and all coefficients



Note: Dots indicate means and whiskers indicate 95% confidence intervals.

Figure D.3.: Study 2: *Quantity* over Altruistic vs. Selfish instantiation and selfish vs. altruistic



Note: Dots indicate means and whiskers indicate 95% confidence intervals.

Figure D.4.: Study 2: $Quantity_{adjusted}$ over Altruistic vs. Selfish instantiation and selfish vs. altruistic

Dep. Var.: $Quantity_{adj.}$	(1)	(2)
Selfish	9.982**	11.397**
	(4.034)	(4.785)
Female		-7.740
		(5.804)
Age		0.022
		(0.251)
UK		-13.055*
		(7.337)
ZA		-4.803
		(8.422)
Altruism		-0.490
		(6.510)
Constant	5.461***	14.496
	(1.246)	(13.022)
R^2	0.132	0.368
Observations	37	37

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.17.: Study 2: OLS regressions with $Quantity_{adjusted}$ as dependent variable and only "altruistic" crowdworkers

Dep. Var.: $Quantity_{adj.}$	(1)	(2)
Altruistic Pref.	-12.420**	-13.318**
	(4.584)	(5.226)
Female		-5.212
		(5.604)
Age		0.194
		(0.201)
UK		2.987
		(8.396)
ZA		5.263
		(5.553)
Constant	17.880***	12.236
	(4.409)	(12.680)
R^2	0.246	0.282
Observations	30	30

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table D.18.: Study 2: OLS regressions with $Quantity_{adjusted}$ as dependent variable and only crowdworkers in the altruistic instantiation

References to Code Repositories, Study Procedures, and Data Sets

The code of the prototypes and the data underlying each of the four studies are available (on request) on the KIT research data repository. Study I has not produced any artifact or data and is therefore not listed here.

Study II

Code: https://git.scc.kit.edu/h-lab/research/haug_saskia_feeasy-crowdfeedback-experiment

Study Data and Procedures: https://radar.kit.edu/radar/de/dataset/zmdRzgZHbtddYlEH

Study III

Code: https://git.scc.kit.edu/h-lab/research/2154_haug_saskia_feeasy-configurator

Study Data and Procedures: https://radar.kit.edu/radar/de/dataset/JloebYgNJdVsfsiO

Study IV

Code: https://git.scc.kit.edu/h-lab/research/haug_saskia_crowdsurfer

Study Data and Procedures: https://radar.kit.edu/radar/de/dataset/oOOnkhdObZSWcOZw

Study V

Code: https://gitlab.kit.edu/kit/iism/h-lab/research/2257_haug_saskia_MyCrowdSurfer

Study Data and Procedures: https://radar.kit.edu/radar/de/dataset/QcDAkdxAOVVJNaWr

List of Publications

Journal Publications

Haug, S.; Benke, I.; and Maedche, A. (2023). Aligning Crowdworker Perspectives and Feedback Outcomes in Crowd-Feedback System Design. *In Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 23.*

Conference Proceedings

Haug, S.; Fischer, D.; Benke, I.; and Maedche, A. (2023). CrowdSurfer: Seamlessly Integrating Crowd-Feedback Tasks into Everyday Internet Surfing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).*

Haug, S.; Sommerrock, S.; Benke, I.; and Maedche, A. (2023). Scalable Design Evaluation for Everyone! Designing Configuration Systems for Crowd-Feedback Request Generation. In *Proceedings of Mensch und Computer 2023 (MuC '23).*

Haug, S.; Ruoff, M.; and Gnewuch, U. (2022). The Impact of Conversational Assistance on the Effective Use of Forecasting Support Systems: A Framed Field Experiment. In *ICIS 2022 Proceedings.*

Haug, S.; and Maedche, A. (2021). Crowd-Feedback in Information Systems Development: A State-of-the-Art Review (2021). In *ICIS 2021 Proceedings.*

Haug, S.; Rietz, T.; and Maedche, A. (2021). Accelerating Deductive Coding of Qualitative Data: An Experimental Study on the Applicability of Crowdsourcing. In *Proceedings of Mensch und Computer 2021 (MuC '21).*

Workshops & Extended Abstracts

Benke, I.; **Haug, S.**; and Maedche, A. (2023). The Human-in-the-loop CrowdSurfer Concept: Providing User-centered AI Support to Crowdworkers for Improved Working Conditions and Task Outcomes. In *Mensch und Computer 2023 - Workshopband.*

Haug, S.; and Maedche, A. (2021). Feeasy: An Interactive Crowd-Feedback System. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21 Adjunct).*

Publications Under Review or in Preparation for Submission

Haug, S.; Benke, I.; Fischer, D.; Walther, S.; Nieken, P.; and Maedche, A. Preference-based Personalization of Casual Microtasking for Crowdworkers. *Working Paper.*

Schloss, D.; **Haug, S.**; and Littwin, E. "Was this Answer Helpful?" – A Taxonomy for the Design of Feedback Mechanisms in Customer Service Chatbots. Under Review at the *European Conference on Information Systems (ECIS) 2024.*

Eidesstattliche Versicherung

gemäß § 13 Abs. 2 Ziff. 3 der Promotionsordnung des Karlsruher Instituts für Technologie für die KIT-Fakultät für Wirtschaftswissenschaften

1. Bei der eingereichten Dissertation zu dem Thema *Human-Centered Crowd Feedback for Information Systems Development* handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, den 13.12.2023

Saskia Haug