
SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
Dissertation



Supporting Users' Influence in Gamification Settings and Game Live-Streams

Dissertation zur Erlangung des Grades des
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

vorgelegt von
(M.Sc.) Pascal Lessel
Saarbrücken
Juli 2018

Date of the colloquium:	November 6, 2018
Dean:	Professor Dr. Sebastian Hack
Reporter:	Professor Dr. Antonio Krüger Professor Dr. Maic Masuch
Chairman of the Examination board:	Professor Dr. Vera Demberg
Scientific Assistant:	Dr. Gerrit Kahl

Notes on Style:

The majority of the work that is presented in this dissertation was done in collaboration with researchers and students. For this reason, the scientific plural “we” is used throughout the thesis. References to web resources (e.g., links to articles) are provided as URLs. Longer URLs have been shortened. Furthermore, most of the studies and systems elaborated on in this thesis targeted a German audience and thus were in German. Images in this thesis depicting these components were translated to English, where necessary.

Acknowledgments

I am in the happy position to thank several people at this very spot in my dissertation. I truly hope that I have not forgotten someone. I apologize if this is the case.

I thank Professor Antonio Krüger, who shares my passion for gaming and allowed me to investigate this topic. Thank you for supervising me, for giving me valuable feedback on my papers and on this dissertation in particular, especially in times when this was not self-evident. Further, I thank Professor Maic Masuch, who agreed to be a reviewer for this thesis. I hope you enjoyed reading it.

I thank Margaret De Lap who not only has proofread this dissertation, but also my publications in the last few years. I am sure that your efforts contributed significantly. Thanks as well to the *Saarbrücken Graduate School of Computer Science* for supporting my work since the preparatory phase. I am convinced that without the offered support, I would not have considered doing a dissertation at all.

I also thank Maximilian Altmeyer, who worked with me as a bachelor's and master's student and now is also working on gamification in his dissertation. I hope you will not curse me at some distant point in the future, for sparking your interest in this topic. Thank you for reading this work and providing valuable comments. I also thank Frederic Kerber and Sönke Knoch for our collaborations in the various *DFKI* projects, especially during *SmartF-IT*. I think I will never forget our business trips to Bünde (and everything connected to it). I also thank all my colleagues (you know who you are!) for providing a fruitful and pleasant working atmosphere and discussions, and of course, the fun we had. I am especially glad that we managed to establish a "regularly meeting" pen & paper role-playing group (before you insist otherwise, once a year is also regularly). Furthermore, I thank all students who worked with me. Sorry for being such a committed and demanding (and sometimes a bit crazy) adviser. I do hope (and I have already several positive examples!) that it has helped you in your further endeavors. Special thanks go to Maximilian Altmeyer (yes, again!), Alexander Vielhauer and Christian Wolff, as the results of our collaborations are now part of this dissertation. Additionally, I thank my friends for their support and help. Maybe I am now going to stop inviting you to experiments (or maybe not...).

Most important to me, I express my deepest gratitude to my parents. They made it possible, through various means, that I could fully concentrate on my studies. As they often stated, they wanted me to have a better starting point in life compared to their youth and early adulthood. And yes, through their support, I can say that this was the case. Thank you.

Abstract

Playing games has long been important to mankind. One reason for this is the associated autonomy, as players can decide on many aspects on their own and can shape the experience. Game-related sub-fields have appeared in Human-Computer Interaction where this autonomy is questionable: in this thesis, we consider gamification and game live-streams and here, we support the users' influence at runtime. We hypothesize that this should affect the perception of autonomy and should lead to positive effects overall. Our contribution is three-fold: first, we investigate crowd-based, self-sustaining systems in which the user's influence directly impacts the outcome of the system's service. We show that users are willing to expend effort in such systems even without additional motivation, but that gamification is still beneficial here. Second, we introduce "bottom-up" gamification, i.e., the idea of self-tailored gamification. Here, users have full control over the gamification used in a system, i.e., they can set it up as they see fit at the system's runtime. Through user studies, we show that this has positive behavioral effects and thus adds to the ongoing efforts to move away from "one-size-fits-all" solutions. Third, we investigate how to make gaming live-streams more interactive, and how viewers perceive this. We also consider shared game control settings in live-streams, in which viewers have full control, and we contribute options to support viewers' self-administration here.

Zusammenfassung

Seit jeher nehmen Spiele im Leben der Menschen eine wichtige Rolle ein. Ein Grund hierfür ist die damit einhergehende Autonomie, mit der Spielende Aspekte des Spielerlebnisses gestalten können. Spiele-bezogene Teilbereiche werden innerhalb der Mensch-Maschine-Interaktion untersucht, bei denen dieser Aspekt jedoch diskutabel ist: In dieser Arbeit betrachten wir Gamification und Spiele Live-Streams und geben Anwendern mehr Einfluss. Wir stellen die Hypothese auf, dass sich dies auf die Autonomie auswirkt und zu positiven Effekten führt. Der Beitrag dieser Dissertation ist dreistufig: Wir untersuchen crowdbasierte, selbsterhaltende Systeme, bei denen die Einflussnahme des Einzelnen sich auf das Systemergebnis auswirkt. Wir zeigen, dass Nutzer aus eigenem Antrieb bereit sind, sich hier einzubringen, der Einfluss von Gamification sich aber förderlich auswirkt. Im zweiten Schritt führen wir "bottom-up" Gamification ein. Hier hat der Nutzer die volle Kontrolle über die Gamification und kann sie nach eigenem Ermessen zur Laufzeit einstellen. An Hand von Nutzerstudien belegen wir daraus resultierende positive Verhaltenseffekte, was die anhaltenden Bemühungen bestärkt, individuelle Gamification-Konzepte anzubieten. Im dritten Schritt untersuchen wir, wie typische Spiele Live-Streams für Zuschauer interaktiver gestaltet werden können. Zudem betrachten wir Fälle, in denen Zuschauer die gemeinsame Kontrolle über ein Spiel ausüben und wie dies technologisch unterstützt werden kann.

List of Publications

Parts of the work presented in this dissertation, including ideas, applications, studies, results, conclusions and other text passages as well as figures and tables have already been published. The following list provides the type of publication, its reference and where it appears in this dissertation:

Full conference papers:

- [5] Maximilian Altmeyer, Pascal Lessel and Antonio Krüger. 2016. Expense Control: A Gamified, Semi-Automated, Crowd-Based Approach for Receipt Capturing. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, 31–42. (appears in Section 3.2)
- [166] Pascal Lessel, Maximilian Altmeyer and Antonio Krüger. 2015. Analysis of Recycling Capabilities of Individuals and Crowds to Encourage and Educate People to Separate Their Garbage Playfully. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 1095–1104. (appears in Section 3.3)
- [169] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff and Antonio Krüger. 2016. “Don’t Whip Me With Your Games”: Investigating “Bottom-Up” Gamification. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2026–2037. (appears in Section 4.2 and Section 4.3)
- [170] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff and Antonio Krüger. 2017. Measuring the Effect of “Bottom-Up” Gamification in a Micro-task Setting. In *Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek '17)*. ACM, 63–72. (appears in Section 4.4)
- [167] Pascal Lessel, Maximilian Altmeyer and Antonio Krüger. 2018. Users As Game Designers: Analyzing Gamification Concepts in a “Bottom-Up” Setting. In *Proceedings of the 22nd International Academic Mindtrek Conference (AcademicMindtrek '18)*. ACM, 1–10. (appears in Section 4.5)
- [171] Pascal Lessel, Michael Mauderer, Christian Wolff and Antonio Krüger. 2017. Let’s Play My Way: Investigating Audience Influence in User-Generated Gaming Live-Streams. In *Proceedings of the 15th ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17)*. ACM, 51–63. (appears in Section 5.2.1, Section 6.2 and Section 6.4)
- [168] Pascal Lessel, Maximilian Altmeyer and Antonio Krüger. 2018. Viewers’ Perception of Elements Used in Game Live-Streams. In *Proceedings of the 22nd International Academic Mindtrek Conference (AcademicMindtrek '18)*. ACM, 59–68. (appears in Section 5.2.2)
- [172] Pascal Lessel, Alexander Vielhauer and Antonio Krüger. 2017. CrowdChess: A System to Investigate Shared Game Control in Live-Streams. In *Proceedings of the 4th Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, 389–400. (appears in Section 6.3)

Short conference papers:

- [173] Pascal Lessel, Alexander Vielhauer and Antonio Krüger. 2017. Expanding Video Game Live-Streams with Enhanced Communication Channels: A Case Study. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 1571–1576. (appears in Section 5.3)

Workshops:

- [165] Pascal Lessel and Maximilian Altmeyer. 2017. Tabletop Game Meets Live-Streaming: Empowering the Audience. In *CHI PLAY '17 Workshop on Augmented Tabletop Games*. 28–32. (appears in Section 5.2.1)

The author of this thesis supervised several bachelor's and master's students. At the time of the submission, six bachelor's theses (BA, with five more currently ongoing), three master's theses (MA, with four more currently ongoing), and one Master Practical Training (MPT) have been finished. The finished works that are relevant in this thesis' context (e.g., having a gamification or live-streaming focus) are enumerated below. Implementations, studies and results that were created and/or described in the student's final delivery during this supervision and which in parts are presented in the above papers (and thus this dissertation) are marked with * and the reference to corresponding paper:

- Maximilian Altmeyer. 2014. Ein gamifizierter und crowd-basierter Ansatz zur Motivation bewusster Mülltrennung*. BA. [166].
- Dominic Gottwalles. 2014. Ein gamifiziertes System zur Verbesserung der Kommunikation und der Übersicht in Großraumdiskotheken. BA.
- Christian Wolff. 2015. Analyzing Crowd Interaction within a Collaborative Gaming Scenario*. BA. [171].
- Christopher Schommer. 2015. To-Do List Applications for Smartphones – Investigating the Current State of the Art and Developing Design Guidelines in a Participatory Design Process. BA.
- Alexander Vielhauer. 2016. Enriched Livestream Experience: Expanding Hearthstone-Livestreams with New Communication Channels*. BA. [173].
- Maximilian Altmeyer. 2016. ExpenseControl – A Gamified, Semi-Automated, Crowd-Based Approach For Receipt Capturing*. MA. [5].
- Sarah Sterz. 2017. The Encouragement of Walking with Gamified Mobile Applications – Development of a Research Prototype. BA.
- Matthias Hennemann. 2017. Twitch Plays Hedgewars: A Case Study of the “Twitch Plays” Phenomenon. MA.
- Alexander Vielhauer. 2017. Development of a Collaborative Chess Game for HCI Studies about Live-Streaming*. MPT. [172].

Contents

1	Introduction	1
1.1	Homo Ludens – Man the Player	1
1.2	Human-Computer Interaction and Games	4
1.2.1	Gamification	6
1.2.2	Game Live-Streams	8
1.3	Problem Statement and Motivation	9
1.4	Research Questions	12
1.5	Contributions to the Field	14
1.6	Thesis Outline	15
2	Background and Related Work	19
2.1	Research Methodology	19
2.2	Supporting Users’ Influence	22
2.2.1	Intrinsic and Extrinsic Motivation	22
2.2.2	Self-Determination Theory	23
2.2.3	The Effects of Choice	24
2.2.4	Games and Self-Determination Theory	27
2.2.5	Users as Designers	28
2.2.6	Summary	31
2.3	Tailoring Gamification	31
2.3.1	Personalization	34
2.3.2	Customization	35
2.3.3	Summary	41
2.4	Crowdsourcing	42
2.4.1	Motivation in Crowdsourcing	43
2.4.2	Image-Based Tasks	46
2.4.3	Aggregation Methods	49
2.4.4	Summary	51
2.5	Live-Streaming	51
2.5.1	Empowering Audience Interactions	54
2.5.2	Shared Game Control	62
2.5.3	Summary	69
3	Gamified Self-Sustaining Systems	71
3.1	Introduction	71
3.2	A Self-Sustaining Household Accounting Book	73
3.2.1	Concept and System Design of <i>ExpenseControl</i>	74
3.2.2	User Study With <i>ExpenseControl</i>	82
3.2.3	Contribution to the Thesis’ Questions	87
3.3	A Self-Sustaining System to Improve Recycling	88
3.3.1	Studying the Wisdom of Crowds in Waste Recycling	90
3.3.2	Concept and System Design of the <i>Trash Game</i>	98
3.3.3	Concept Evaluation of the <i>Trash Game</i>	104
3.3.4	Contribution to the Thesis’ Questions	106
3.4	Summary	107

4	Self-Tailored Gamification	109
4.1	Introduction	109
4.2	“Bottom-Up” Gamification	111
4.2.1	“Bottom-Up” Game Elements	112
4.2.2	A Priori User Assessment of “Bottom-Up” Gamification	114
4.3	A “Bottom-Up” Task Management Application	120
4.3.1	Concept and System Design of <i>BU-ToDo</i>	121
4.3.2	User Study with <i>BU-ToDo</i>	124
4.3.3	Contribution to the Thesis’ Questions	129
4.4	“Bottom-Up” Gamification in a Microtask Setting	130
4.4.1	Concept and System Design of the <i>BU-Microtasks Platform</i>	131
4.4.2	User Study with the <i>BU-Microtasks Platform</i>	134
4.4.3	Contribution to the Thesis’ Questions	142
4.5	Can Users Create Suitable Gamification Concepts?	143
4.5.1	User Study	143
4.5.2	Contribution to the Thesis’ Questions	159
4.6	Summary	159
5	Interactivity in Game Live-Streams	161
5.1	Introduction	161
5.2	Analysis of Interactivity in Game Live-Streams	163
5.2.1	Interactivity Today	163
5.2.2	Viewers’ Perception of Interactivity	169
5.2.3	Contribution to the Thesis’ Questions	185
5.3	Enhancing Interactivity in Game Live-Streams	185
5.3.1	Concept and System Design of <i>Helpstone</i>	186
5.3.2	User Study with <i>Helpstone</i>	191
5.3.3	Contribution to the Thesis’ Questions	196
5.4	Summary	196
6	Shared Game Control in Live-Streams	199
6.1	Introduction	199
6.2	Perception of a Shared Game Control Setting	200
6.2.1	Concept and System Design of <i>TPP++</i>	202
6.2.2	User Study with <i>TPP++</i>	204
6.2.3	Contribution to the Thesis’ Questions	210
6.3	Effectiveness in a Shared Game Control Setting	211
6.3.1	Concept and System Design of <i>CrowdChess</i>	211
6.3.2	User Study with <i>CrowdChess</i>	217
6.3.3	Contribution to the Thesis’ Questions	226
6.4	Summary	226
7	General Conclusions	229
7.1	Summary	229
7.2	Major Contributions	230
7.3	Future Work	232
	List of Figures	235
	List of Tables	237
	Bibliography	239

Chapter 1

Introduction

In this chapter, we give a short overview on the role of gaming through history, the impact it has today and how this led to gaming-related sub-fields that are considered within Human-Computer Interaction (HCI). We put special emphasis on the gamification and game live-streams domains as this thesis contributes to these sub-fields. We motivate the problem this work targets by discussing which options these offer (or fail to offer) in relation to influence individual users can exert. Based on this, we present the questions that guided the research done during the dissertation. This chapter concludes by describing the contributions we make, by providing an overview on the thesis' structure and how the chapters relate to the research questions.

1.1 Homo Ludens – Man the Player

Playing games has long already been a topic for mankind, and it still is, one that even “*suberves culture*” ([123], p. 9). Evolutionarily, this is unsurprising as humans possess an inherent ludic drive [152] and even before children learn to speak, they can understand their environment playfully, helping them to learn [113] (e.g., in language acquisition [243]). Even after childhood, it remains active lifelong [152]. Huizinga sees the aspect of playing as so important that he characterized humans with the term *Homo ludens* (Man the Player) in contrast to *Homo faber* (Man the Maker) [123], which, anthropologically, are subordinates of *Homo sapiens*. Traces of toys and games were found already in the new stone age (*Neolithic*, beginning about 10,200 BC) and it is assumed that even here, games were played that did not require any toys [9]. Other reports show that dedicated board games (e.g., such as *Mehen* or the *Royal Game of Ur*) were found that dated



Figure 1.1: Chess as an example of a game impacted by technology. Left: An analog chess game. Right: Online platform for playing chess (screenshot was taken from <https://lichess.org> in July 2018).

back to 3100¹ to 2600 BC². Games were used to mimic cultural activities: games requiring physical skill were seen in relation to combat, chance-based game in relation to religion and strategic games simulated war [35, 250]. An example of the latter is *chaturanga* which can be traced back to the sixth century in India [55]. *Chaturanga* can be seen as a predecessor of *chess*, which is played today, not only as a co-located board game, but also via online platforms, showing that games persist and that technology also has an impact on gaming (see Figure 1.1).

A definition of play is “a voluntary activity or occupation executed within certain fixed limits of time and place, according to rules freely accepted but absolutely binding, having its aim in itself and accompanied by a feeling of tension, joy and the consciousness that it is ‘different’ from ‘ordinary life’” ([123], p. 28) and “play to order is no longer play” (ibid., p. 7). In later work, games and play are differentiated into *paidia* and *ludus* [33]. While *paidia* denotes an improvised, free form of behavior and relates to playing and toys (e.g., children playing with puppets), *ludus* denotes a structured activity with rules and a clear goal relating to gaming and games (e.g., playing a board game) [64]. In this thesis, we focus on the latter, but also make use of *paidia* aspects by providing users with more freedom. Schell [264] highlights that defining play and games is not easy and “that the ideas these terms represent do not have clear definitions even after the thousands of years we’ve been thinking and talking about them” (p. 25). Thus, unsurprisingly, there are many definitions for games “which [have] always been at the centre of multiple perspectives and conflicting approaches” ([35], p. 55). What is derived from them, though, is that games (amongst other aspects) are “entered willfully”, “engage players”, “are interactive” and “have rules” (all [264], p. 34).

Today, the range of existing game genres (and the instances in these genres) can be considered large [104, 264]: activity-based group games (e.g., *volleyball*), board games (e.g., *Backgammon*), card games (e.g., *Rummy*), pen and paper role-playing games (e.g., *Dungeons and Dragons*), mobile games (e.g., *Angry Birds*) or video games (e.g., *StarCraft*) are just a few examples that have even more

¹ Ancient Games: *Mehen*, <https://goo.gl/QDNU6H> (last accessed: 2018-07-07)

² BBC: *Assyrian guardian figure*, <https://goo.gl/yjNkPt> (last accessed: 2018-07-07)



Figure 1.2: A historical comparison of games. Left: *Pong* (Created by Atari, 1972). Right: *The Witcher 3* (Created by CD Projekt Red, 2015; screenshot was taken from press kit, <https://goo.gl/M3Tsq4> (last accessed: 2018-07-07)).

subcategories. In these cases (similar to the *chess* example above) technology altered not only the available categories, but also how we play and perceive games [17]. Video games appear to be most affected by the technology available: while first reports of machines allowing one to play games date back to 1940, the start of the commercial impact of video games begun in the seventies³. While the first video games had simple game principles and graphics, today these are rich in graphics (see Figure 1.2 as a comparison of *Pong* (1972) and *The Witcher 3* (2015)) and offer complex game dynamics (the time to play through *The Witcher 3* is said to be 70+ hours, when all optional aspects are also completed⁴).

The impact of video games can be considered high: the *Entertainment Software Association* (ESA) has released data⁵ showing the economical impact of video gaming, as the US game company industry's value added to US GDP was more than \$11.7 billion in 2016. It thus generates more revenue than other entertainment industries [211]. Based on more than 4,000 American households, *ESA* also reported that 64% of these own a device that is used to play video games and considering the gamers, 45% are women. Video gaming is also not bound to an age group, relating back to the ludic drive and desire to play: although more younger people claimed to game, the numbers of older people that are still playing is notable (male: <18 years: 17%, 18–35: 16%, 36–49: 12%, >50: 11% and female: <18 years: 11%, 18–35: 13%, 36–49: 8%, >50: 12%). The situation is similar in Germany: a 2017 *Bitkom* study⁶ (n=1,192) found that 46% of the male and 41% of the female participants reported playing video games at least occasionally and that gaming is a relevant topic also across the different age groups (14-29 years: 74%, 30–49: 63%, 50–64: 24%, >64: 12%).

³ Museum of Play: *Video Game History Timeline*, <https://goo.gl/pkXbnJ> (last accessed: 2018-07-07)

⁴ Forbes: 'The Witcher 3' And Game Length As A Mountain To Climb, <https://goo.gl/jqnA2y> (last accessed: 2018-07-07)

⁵ ESA: *2018 Essential Facts About the Computer and Video Game Industry*, <https://goo.gl/kKCEvy> (last accessed: 2018-07-07)

⁶ Bitkom: *Mobil und vernetzt: Die Gaming-Trends 2017*, <https://goo.gl/WNNhKz> (last accessed: 2018-07-07)

The relevancy of video gaming is also acknowledged by scholars, highlighting that video games are a next step in the evolution of the game culture: Deterding et al. used the term “*ludification of culture*” and reference media scholars view “*that video games have become a cultural medium... on a par with literature, movies or television in earlier generations. Technologies, tropes, references and metaphors, mindsets and practices flowing from games increasingly suffuse society and everyday life*” (both [64], p. 10). Connolly et al. also stated that “*over the last 40 years computer games have increasingly replaced more traditional games as leisure activities and have had a transformational impact on how we spend our leisure time*” ([49], p. 661). Although the learning aspect of games [84] mentioned above might not be predominant for adults anymore, they still impact individual aspects (e.g., cognitive skill improvements [49]). Additionally, other investigated attributes of video games, such as their entertainment value [62, 104, 211], the social aspects of playing with others [104, 164, 297] and their power to trigger behavior change [96, 227] are all reasons why people spend billions of hours playing games [314]. Both the inherent human desire to play and the apparent success of games are drivers for the thesis’ questions, as we will describe below.

1.2 Human-Computer Interaction and Games

Human-Computer Interaction (HCI) is a “*discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them*” ([111], p. 5). According to [65], *Human* in HCI means individual users or a group of users. *Computer* covers any technology ranging from embedded systems to large-scale computer systems and may include non-computerized parts (e.g., other people). *Interaction* means any communication between a user and computer in order to achieve something.

In the 2000s, games started to receive attention in HCI as researchers began to study design and experiences of video games [64], today termed *games user research* [211] or *game studies* [8]. Besides producing a large body of knowledge to date [104], the relevancy for gaming in HCI was also formulated in [213]: “*It is now more important than ever to study the human impact of video games. There are now more games papers submitted to CHI than, for example, health-related papers. Studying video games is a serious and important area of research at CHI that continues to grow as games pervade into many areas of research*” (p. 1089). Methods commonly used in HCI research are applied in game research as well, but new methods and instruments, particularly relevant for this context, have also been developed. An example of the latter is the *Player Experience of Need Satisfaction* questionnaire to assess the players’ enjoyment [249], but existing questionnaires were also adapted to better fit the (video) game context, such as the *User Engagement Scale* [327]. Besides research on games in particular, another development can be seen: Deterding et al. [64] emphasized that in HCI, available game components and game elements have often been “re-purposed” in a different context. One

example of this is the usage of the *Microsoft Kinect* sensor, an input device for the *Xbox* gaming console, which has often been used in non-game scenarios (e.g., [202]). The idea to utilize ideas from games in different contexts was not new to HCI. After the first success of video games in the past, researchers had already started to follow the idea to design user interfaces that are enjoyable by utilizing aspects from games to analyze motivational affordance and to learn how to create pleasurable products [64]. Today, in HCI multiple game-related sub-fields have appeared (see below). Following our above argumentation, this is unsurprising given the popularity of games, the idea to harness the *Homo ludens*, the positive impact games have on individual attributes and that “*games, after all, were known to be fun and engaging by nature, a quality most software applications were still struggling with*” ([211], p. 50). Examples of such sub-fields are:

- **Serious games:** This line of research considers games that do not have the primary purpose of being entertaining, but, for example, are instead tools for helping people to learn by framing the learning task in a game scenario [1, 49].
- **Games with a purpose:** Also known as human computation games. These are games that are enjoyable, but by playing them, players solve large-scale problems in parallel [314]. An example is the *ESP game*, a game about generating image labels [315] (see Section 2.4.2).
- **Gamification:** Here it is investigated how elements and mechanics known from games can be used in contexts that have nothing to do with gaming in general to make these systems or activities more fun or engaging [64].

But the gaming culture is also a topic considered in HCI, for example:

- **eSports research:** Many multiplayer games offer competitive settings that are similarly appealing as football or basketball matches are for spectators. The attributes of competitive computer gaming is analyzed in this line of research [80, 320].
- **Game live-stream research:** The goal is to understand the current practices around game live-streams (i.e., people show how they play games, which attracts large amounts of viewers), what motivates viewers to watch and streamers to stream and how interactivity can be supported for them [108].

Considering the thesis context, these developments are important: while these sub-fields originated from games in general, some of them restrict their users by not focusing on core attributes important to games such as voluntariness or interactivity (as we will discuss below). As we aim to empower users in the gamification and game live-stream sub-fields in this respect, we will introduce these areas more specifically next, before we will detail the thesis’ problem statement and the research questions.

1.2.1 Gamification

Gamification originates from the business and marketing sectors [269]. Even though the method and term itself are not free from criticism [61, 269], as there are no clearly defined boundaries for example, they remained [324]. Considering the increasing numbers of publications [106] gamification has become a relevant topic in academia and is discussed in the media⁷. There are several definitions of gamification [124, 324], but for this thesis we consider Deterding et al.'s commonly used definition (with more than 3700 citations (*Google Scholar*) in July 2018) that it “*is the use of game design elements in non-game contexts*” ([64], p. 10).

Following this definition, gamification introduces characteristic mechanics and/or elements known from either analog or digital games (see [106, 134, 269] for an overview) to areas, across domains, that are usually not associated with gaming in general (e.g., learning [27, 106] or physical activity [272, 338]). In contrast to the above-mentioned serious games or games with a purpose, where a game is presented with other components (such as learning elements), in gamification, it is the other way round. Here, game elements are added to an existing system [205, 234]. The non-game focus of the system is thereby unchanged and the game elements are only a byproduct. The goal is to make tasks more fun, engaging and rewarding [24, 269, 303], but it can also be used to increase aspects such as the efficiency in task solving [199]. In addition, gamification is used to influence behavior playfully and thus can be seen as strategy in persuasive technologies [27]. Research considers how to design successful gamification approaches to achieve these goals (see [62] for an overview) and positive effects of gamified intervention are reported throughout research. For example, Shameli et al. [272] analyzed a ten-month dataset of an application that allows users to create competitions with other users of the form “*Who does the most steps in the next X days?*”. This is an example of a gamification approach that aims at influencing behavior playfully. The authors found that it motivated people to do 23% more steps. Literature reviews done by Hamari and Tuunanen [107] and Seaborn and Fels [269] show further positive results of gamification on a larger scale. Gamification is also frequently used in non-academic areas. We selected two popular web pages to give examples of how gamification is applied:

- The business and employment-oriented social network *LinkedIn*⁸ utilizes gamification concepts to motivate users to add more information to their profiles (see Figure 1.3). Several *game elements* are visible (denoted with black numbered circles in the figure): (1) the colored progress bar shows the profile completeness relating to the elements *feedback and progression*; (2) a *level* indicating the informative value of the profile; (3) three *milestones* are shown on the progress bar that are accompanied with *badges* when reached; and (4) *motivational messages and hints* are given.

⁷ For example: The New York Times: *All the World's a Game, and Business Is a Player*, <https://goo.gl/hVhyjS> (last accessed: 2018-07-07)

⁸ <https://www.linkedin.com> (last accessed: 2018-07-07)

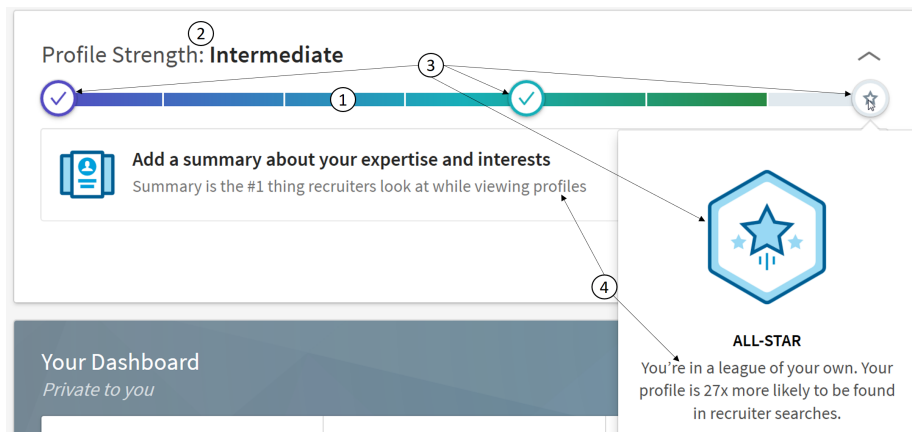


Figure 1.3: Profile view excerpt from the social network *LinkedIn* (screenshot was taken from <https://www.linkedin.com> in July 2018).

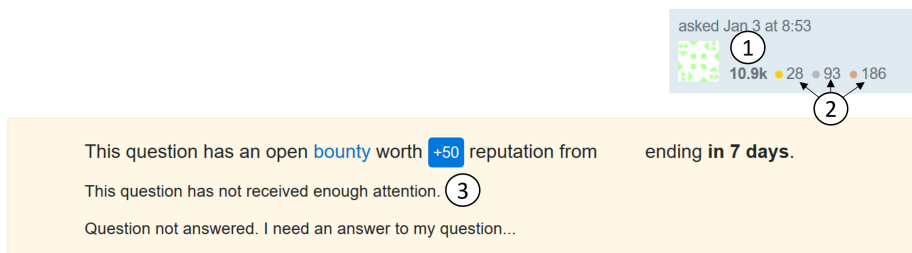


Figure 1.4: Excerpt from the Q&A site *Stack Overflow* (screenshot was taken from <https://stackoverflow.com> in July 2018).

- *Stack Overflow*⁹ is a question and answer (Q&A) page for computer programming. It offers features that aim at improving the quality of the answers, such as the chance for users to upvote given answers to identify the most helpful one. Besides the inherent utility of these features, their usage is further motivated through gamification. Figure 1.4 shows an excerpt of an area that is displayed on every user's post: (1) a *point* score represents the user reputation (which *unlocks* page features) that can be increased by, for example, receiving upvotes from other users; (2) a representation of how many *badges* of different categories (bronze, silver, gold) a user has received, which can lead to *social recognition* and potentially *competition*. Badges are awarded for different behaviors on the platform. For example, the bronze badge *Scholar* is rewarded when a user asks a question and accepts an answer as correct; and (3) points receive *meaning*: they can be spent as bounty to make answering the question more appealing to others.

⁹<https://stackoverflow.com> (last accessed: 2018-07-07)

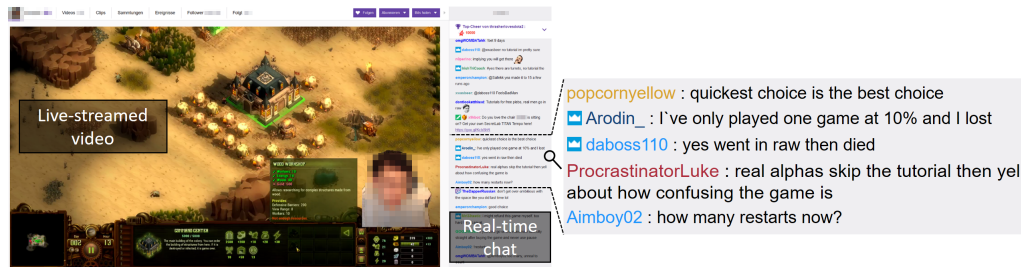


Figure 1.5: Example live-stream on *Twitch* (screenshot was taken from <https://www.twitch.tv> in July 2018). Part of the chat enlarged for readability.

1.2.2 Game Live-Streams

Advances in technology have created settings where interactions between performers and their audiences can be altered, enhanced and mediated through technology [36, 37]. With the advent of faster Internet connections, a new context has appeared. Ordinary people can become performers on their own, and it is easy to consume such performances [180]. An example is the production of user-generated video content [311] live-streamed over the Internet: people (called streamers) show how they cook, dance or mix music [273], or how they experience events [100]; they talk about their personal lives [146], or show how they do programming [99] or how they play analog or digital games [108].

The latter is a particularly successful form of live-streaming and has a large streamer and viewer base today [108, 279] with streams attracting 10,000 or more viewers in parallel [60]. This is not surprising, considering that, originally, sports games such as football were designed for their players but have also attracted millions of spectators over the years; the same seems true for these live-streamed games [214, 293]. In addition, researchers also discussed that spectatorship in games has always been central to gaming [294] and is increasingly regarded as social activity [67]. *Twitch* (owned by *Amazon*) is the most popular page that focuses on live-streaming, ranking (in July 2018) at place 14/34/40 of the top pages in the US/Germany/worldwide¹⁰. *Twitch* requires streamers to mainly broadcast content in relation to analog/digital games or the gaming culture. Typical live-streams include eSports matches, playthroughs of games in which streamers play the game and entertain their audience while doing this (so-called *Let's Plays*) and games in which streamers show their skill (e.g., beating a game as fast as possible, so called *Speedruns*) [278, 279]. Streamers often integrate a webcam view of themselves in their performance [108] (see Figure 1.5). More than two million unique broadcasters performed on *Twitch* in 2017 and viewers watched 355 billion minutes of live-streams¹¹. *Mixer* (owned by *Microsoft*) and *YouTube Gaming* (owned by *Google*) are competitors to *Twitch*, also focusing on live-streamed games, showing the industry's interest in this experience.

¹⁰ According to the *Alexa* traffic rank, <https://www.alexa.com> (last accessed: 2018-07-07)

¹¹ *Twitch: 2017 Year in Review*, <https://goo.gl/3wy6HM> (last accessed: 2018-07-07)

Consuming live-streams is not only a passive activity [83]. Live-streaming platforms today provide a text-based chat as an interactive channel (see Figure 1.5), allowing the streamers to directly react to their audience, to integrate viewers into the performance and allow them to influence how it progresses [100, 292]. Especially these aspects make this context relevant for this thesis, as this can also be used to, for example, impact the game directly [271]. Through the chat, viewers of live-streams can also communicate with each other [109], talk about the performance [100] or socialize around the streamer [83]. This establishes a sense of community that affects how much people watch streams and whether they become regulars [108]. It was shown that the appeal of game live-streams originates from these interactive and social components [108, 278]. In this sense, live-streaming can be seen as successor of *Social TV*, which had the goal to make TV watching more interactive and social [37, 86].

1.3 Problem Statement and Motivation

In Section 1.1, the voluntary nature of games and play was highlighted. In the first place, this means that one can always decide whether or not one wants to play or take part in gaming. Deterding stresses that games are “*something we can choose to do and cease doing – it satisfies our need for autonomy... playing games... is an activity we typically feel we do following our own interests, where we decide what to play, when, how and how long, with no social or material pressures or consequences affixed*” ([61], p. 309). This viewpoint can also be extended to aspects such as being able to adapt the game experience to be a better fit to one’s (or the group’s) needs [17]. While this appears easy in play scenarios (i.e., *paidia*), it might also be possible in game scenarios (i.e., *ludus*), as, for example, in board game contexts, rules can easily be adapted to be a better fit [17]. In digital games the situation becomes more difficult, as for a user it is not as easy to adapt the game foundations without being a programmer, having games that allow for this and/or having access to game modifications [237]. Often, however, through aspects such as in-game decisions, the game’s difficulty and other game aspects that can be adjusted [286], choices are even offered in such games. By considering the following thought experiment, it becomes obvious that games can also be used “wrongly”:

Assume that a superior at work decides that all employees have to play a certain game (that was selected by him) during their break to relieve stress and be entertained, because he had read some articles about the benefits of games in general. As he is a player himself, he knows that games are fun, so he thinks that this will be a good idea. As the superior is convinced that his selected game is perfect, he also decides that the game has to be played every day at least once. He also thinks to know best which game settings have to be used to induce fun and relief. In the following weeks, it becomes obvious that he has failed; the employees are not entertained and dislike his idea.

From this thought experiment three factors can be derived for this failure: first, the superior decided that the employees need to play a game. It is obvious that the voluntary aspect of playing games was not considered here, which was shown to be problematic, especially in the work context [203]. Even if the game might have attracted some employees, the knowledge that they need to play it every day counteracts the freedom of choice that is inherent when it comes to play. Research in relation to coercion also reveals the negative impacts of a general lack of freedom of choice [58, 78]. It might even be the case that the employees now perceived the game as part of work [220]. Second, the superior chose the game, following a “one-size-fits-all” solution, i.e., everyone should play the same game. Besides the fact that this also restricts the freedom of choice (as he had selected the game), it cannot be assumed that everyone likes a game to the same degree, as there are individual differences [110, 226, 306]. In addition, the game setting was also decided by him, so this issue cannot be mitigated by the actual players. Third, although the superior had good intentions for the employees, he did not consider their opinions. This means that although he wanted to deploy something *for* them, they were not able to *contribute* to it in the first place, nor could they consent to it [203]. The failure can also be explained by the *Self-Determination Theory* [256] (which we discuss in more detail in Section 2.2.2) and its consideration of the autonomy need (the feeling of acting under one’s own volition). In the thought experiment, autonomy was neglected for the employees (only the superior acted from an autonomous point of view). While the experiment considered introducing a (full) game to the work context, we see similar issues in the introduced game-related sub-field, as the voluntary nature of gaming is neglected and/or users have only a low impact at runtime:

Gamification: Gamification is also used in the work context (e.g. [149, 153, 203]). The goals of the gamified intervention here can be to motivate employees to do tasks more efficiently which has led to criticism in the past. For example, laundry workers at *Disneyland* in Anaheim saw the introduction of leaderboards comparing their speed as an “*electronic whip*”¹²; a term which has been used in the following by researchers alike (e.g., [61, 211]). Another introduced term in this context is “*exploitationware*”¹³ referring to the issue that employees are exploited in the guise of a game. Thus, the usage of gamification today shows strong parallels to our thought experiment above. In this context, again, it is visible that the voluntary nature of games and play was not considered for the actual users of the system, i.e., some other group of people has decided that gamification should be used and how [61]. Mollick and Rothbard [203] termed this as manager-imposed “*mandatory fun*” (p. 4). But even when gamification is not used in a work context the situation is similar. Considering the *Stack Overflow* and *LinkedIn* examples above, users cannot adjust the game elements to their needs, as the solutions presented are “one-size-fits-all”. The current body of

¹²Forbes: *Disneyland Uses ‘Electronic Whip’ on Employees*,
<https://goo.gl/tUoMJC> (last accessed: 2018-07-07)

¹³Gamasutra: *Persuasive Games: Exploitationware*,
<https://goo.gl/CqVj4j> (last accessed: 2018-07-07)

research shows that there are individual differences in the perception of game elements (e.g., [306]) and that not respecting these is not optimal (e.g., [24, 226]). Thus, when users are not satisfied with the gamified intervention in the examples, they only have the options to abandon the services or to ignore the gamified interventions on the pages (both might not be possible in a work context). This, on the other hand, does not fulfill the goals for why gamification is used on the first place there. Overall, this underlines that offering more freedom of choice at runtime would be reasonable to provide users with the autonomy to influence the gamification of the system they use.

Game Live-Streams: Considering the situation in game live-streams, the streamer has high autonomy and influence (as he or she can decide what happens in his or her stream). The situation for viewers is different. As described above, integrating viewers into the experience is a crucial aspect in game live-streams, but here, how many options the viewers actually will receive depends on the streamer. Even if the streamer is open to allow for more viewers' autonomy, there are technological limitations in the available options. For example, the chat as the primary existing interaction channel has shortcomings: the range of possible interactions can be considered as limited [271]; in large streams, streamers and viewers suffer from information overload [108]; and the communication changes in situations where many viewers are active in parallel [221]. Although authors such as Ford et al. [79] reasoned that interactions are still possible and that the chat remains a successful communication space, other authors compare the situation to the "*roar of a crowd in a stadium*" ([108], p. 1321) in which interactions are problematic [108, 221]. It is concluded that chat as an interaction channel is not sufficient and that alternatives are necessary to facilitate social options and interactivity [121, 277]. Another issue is that interactive options not only cover functions that change something for the user (e.g., deciding for oneself which camera perspective is most interesting to watch), but can also impact all other viewers (e.g., altering the game the streamer is currently playing by making it more difficult). This raises questions about how enhanced influence options can be granted in such group settings. In addition, explorations have shown that there are also individual motivations for why game live-streams are consumed [41, 112]. Thus, allowing individuals to alter the experience to their needs might be beneficial as well.

Summing up, users have only a small impact on gamification and game live-streams at runtime. We see that the freedom of choice and influence options for the consumer of these game-related experiences is limited. As the mentioned related work shows, it would be beneficial to empower users. This would not only strengthen the relationship to games and play in general, but would also make it possible to mitigate issues. If users received more interaction possibilities and/or options to customize the systems, based on the argumentation so far, this could be expected to improve the experiences. It is an open question how users would perceive higher autonomy here and how it could be empowered in gamification and game live-streams. This thesis will contribute to both aspects.

1.4 Research Questions

Following the previous considerations, the question that guided this work is:

Can we provide users with more influence options at the runtime of gamified systems and game live-streams, to provide them with more autonomy, similar to games and play, where these sub-fields originated from?

To make this question more manageable, we have broken down the overall question into the following main research questions to which this thesis contributes:

- RQ1** How can we develop gamified services in which user contributions at runtime are important for the system's outcome?
- RQ2** How can we provide users with the option to define their own, individual, gamification intervention at the runtime of a system?
- RQ3** How can we empower the group of viewers in game live-stream settings to have an impact on the stream while it is running?

With **RQ1**, we investigate systems that can only provide a reasonable result because their users are expending effort to improve the service itself. We call these self-sustaining systems and will analyze how users perceive and how to realize them, whether adding gamification will be beneficial for motivation and whether the participation in such systems also implicitly impacts the users. By targeting this research question, we learn what happens when a user has fundamental influence on the general outcome of a gamified system at runtime. With **RQ2** we investigate whether users actually want to have the option to alter the motivational aspects of a system at runtime, how to allow them to do this and whether they are capable of using the options in such a way that the gamification still has positive effects on fun or engagement. We call the concept of self-tailored gamification at runtime "bottom-up" gamification. By targeting this research question, we learn what happens when a user has a fundamental impact on the motivational aspects of a system. Finally, with **RQ3**, we investigate whether users want to have more influence options in game live-streams, how we can realize improved interaction channels and how we can mediate individual contributions and support self-administration in group scenarios in this context. By targeting this research question, we learn what happens when a user has an impact on other users (i.e., the streamer and other viewers) in a system.

In all three considered cases, the increased user influence alters the experience users have with a system, i.e., there is a reciprocity effect: in self-sustaining systems the quality of the offered service changes; the user influence on gamification alters the motivational impact of the system; and by exerting influence on the game live-stream, the stream, and thus the overall experience, might change.

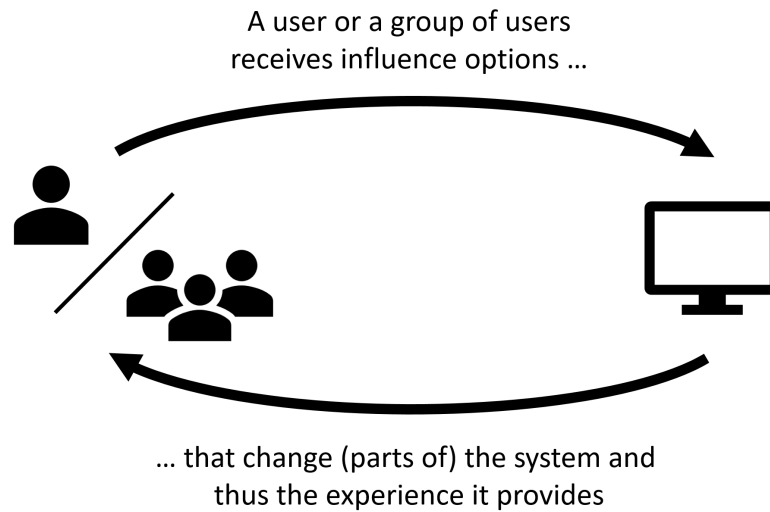


Figure 1.6: General schematic of reciprocity of increased user influence and how users experience the system. This schematic will be instantiated throughout the chapters of this thesis.

Moreover, by exploring these research questions, we will also investigate different user constellations: in self-tailored gamification every individual can decide for themselves and, for example, are only compared to others when they want to be. In the self-sustaining scenarios the user group is loosely coupled (e.g., by, for example the knowledge that everyone can contribute to the system) and, in contrast, the live-streaming context offers a tight group coupling as, depending on the individual impact, the experience might change for everyone. Figure 1.6 shows an abstract visualization of the reciprocity and the individual/group setting which is instantiated throughout the different chapters.

From a methodological point of view (see Section 2.1 for a detailed discussion on the methods used in this thesis), we consider the perception of increased user autonomy in different contexts by analyzing these (e.g., how users perceive interactive options in live-streams) and conceptualizing and realizing prototypical systems that provide users with more influence options (e.g., the development of a system offering more interactive options in a game live-stream). We conduct user studies to gain insights into the actual usage of the developed systems and which effects they have on the users and the systems themselves (e.g., evaluation using this system to learn how the interactive options are used). Specifically, following this, we have formulated goals that add to corresponding research questions. Figure 1.7 shows an overview of these goals. We motivate and elaborate on the different goals throughout the chapters in more detail.

The thesis provides theoretical, design and engineering contributions that will be elaborated on in particular in Chapter 7, after we have presented all parts of this work. In the next section, we already describe the contributions this thesis makes to the different research fields in HCI.



Figure 1.7: Overview of goals that we targeted in the thesis in relation to the research questions. The goals will be motivated and elaborated on in the upcoming chapters.

1.5 Contributions to the Field

Based on the research questions, this thesis contributes to several fields in HCI. Primary, it advances the understanding of gamification and game live-streams research. For gamification, it reveals whether users want to have more autonomy, and investigates how self-tailored gamified systems can be designed, what these need to offer, how they are perceived by users and which qualitatively and quantitatively measurable effects such systems exert. In this sense, it also contributes to the ongoing efforts in designing gamification systems that do not follow a “one-size-fits-all” approach but are adapted to the users based on customization. For game live-streams, the thesis reveals viewers’ needs and requirements to-

wards interactivity and how much impact they want to exert on the course of live-streams. It provides insights on attributes of streamer-audience interaction and contributes aspects on how systems should look to allow a group of people to interact effectively in this context. Secondly, this thesis also contributes to the areas of crowdsourcing (i.e., the idea that a group of people solve specific tasks that are not easily solvable by a computer; see Section 2.4), computer-mediated communication (i.e., how communication can be supported through communication tools [301]) and games user research. It is relevant to crowdsourcing, as the consideration of self-sustaining systems provides insights into their design. Furthermore, many questions in the group setting of live-streaming are also of relevance for crowdsourcing, for example how to mediate individual contributions. As gamification is also often utilized in crowdsourcing [206, 269], the findings for self-tailored gamification are applicable here as well. As live-streaming settings offer many-to-many (viewers amongst themselves) and many-to-one (viewers to streamer) interaction and communication situations that need to be considered to allow for interactivity, we also contribute to computer-mediated communication. Finally, the thesis' outcomes are also of relevance for games user research, as the results in the live-streaming context provide insights on how games that have the goal to empower streamer-audience interaction could be improved.

Non-goals of this work: This thesis only focuses on the game-related areas of gamification and game live-streams and does not consider other game-related sub-fields nor systems that have nothing to do with games or gaming. The former is a consequence of the scope that can be handled within a dissertation. The latter was assessed as less interesting and relevant for this thesis, as it would have shifted away from *Homo ludens*, from the inherent human ludic drive and the argumentation for why games are motivating. This thesis will also not consider games in particular, i.e., we will not consider how to improve user autonomy in typical single- or multiplayer games (although the shared game control settings in Chapter 6 might add to this). Today, (video) games already offer autonomy aspects (e.g., the user's option to customize the game's avatar or to adjust the difficulty level for the game [286]) and such existing mechanisms have already been considered scientifically (see for example [143]). Although providing the players with even more (new) autonomy options in these contexts might be interesting to study, that is not a focus of this thesis.

1.6 Thesis Outline

The remainder of this thesis is structured as follows (see also Figure 1.8): in Chapter 2, we give more details on our research methodology, elaborate on the background and theoretical foundations of this thesis, and present related work and open questions in gamification, (gamified) crowdsourcing, and (game) live-streams. Chapter 3 describes our systems, *ExpenseControl* and the *Trash Game*, two game-based crowdsourcing systems in which the users' contributions at

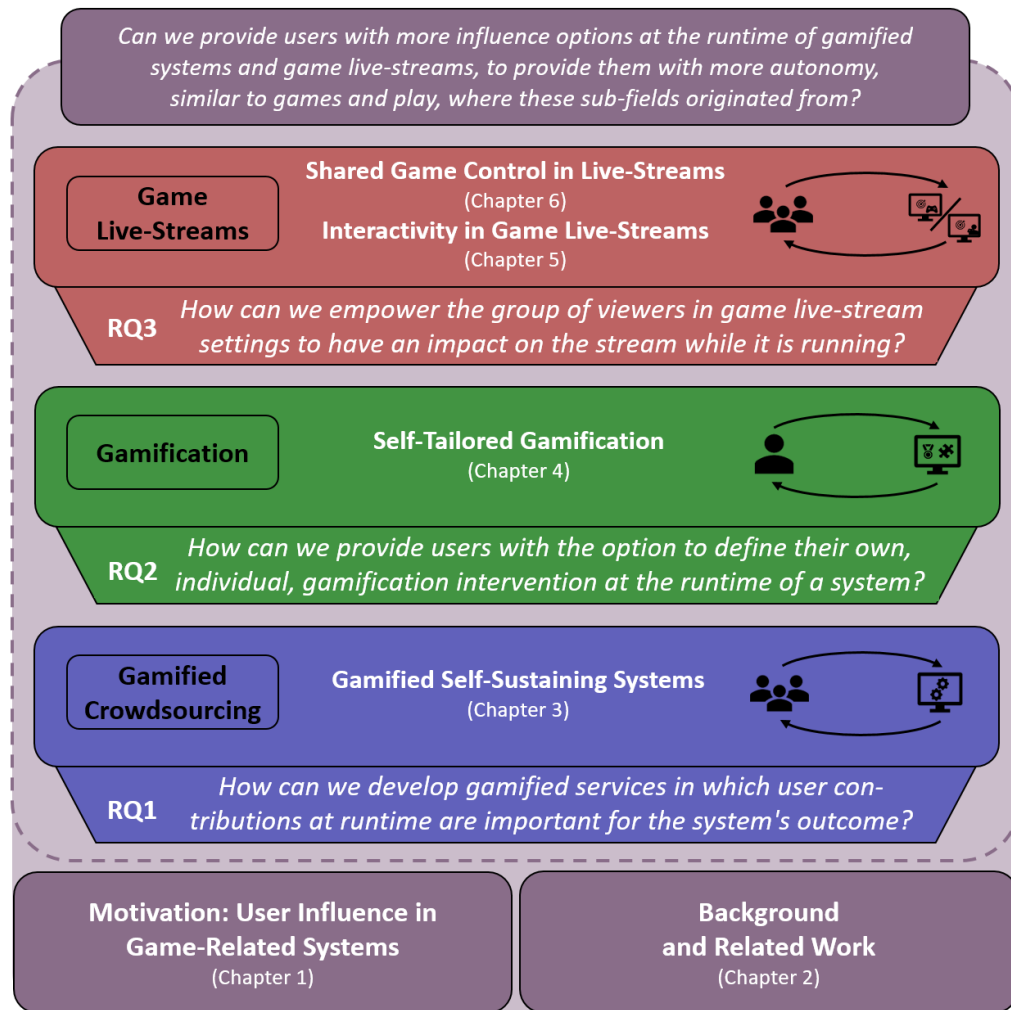


Figure 1.8: Structure of this thesis with chapters and relations to the corresponding main research fields and the instantiated reciprocity schematic, that will be explained in more detail in the respective chapters.

runtime improve the service the systems offer, thus representing self-sustaining systems. We describe the systems' design and present studies that revealed that self-sustaining systems work, that users can be further motivated by gamification in these settings and that implicit effects can be exerted by such systems. Overall, these results add to **RQ1**. Chapter 4 introduces our idea of self-tailored gamification, i.e., the option that users can decide whether they want to use gamification in a system and if they want, can fully customize how they will use it at the system's runtime. We present a study that assessed user expectations without an actual prototype, describe a task management application (*BU-ToDo*) and a microtask platform (*BU-Microtasks Platform*) implementing self-tailored gamification and studies that we conducted with these systems. Through these, we were able to show that the idea of self-tailored gamification is reasonable and

that it exerts positive, qualitatively and quantitatively measurable effects. The chapter closes with a first consideration of where these effects originate. Overall, these results add to **RQ2**. Chapter 5 and Chapter 6 focus on game live-streams. In Chapter 5, we describe a study on recent live-streams, to learn about the influence viewers have today in game live-streams. Additionally, we present the results of an online study in which we assessed viewers' expectations and needs for interactivity in such streams. Based on this, we created *Helpstone*, a system that provides extended interaction channels to the live-streaming experience. With a study in an "in the wild" context, we showed that more influence options and enhanced interaction channels are widely used and are appreciated in live-streaming. This, even for viewers that do not want to actively exert influence, as long as the streamer is able to orchestrate the viewers. While Chapter 5 focuses on streams in which a streamer is present, Chapter 6 focuses on streams in which a streamer is absent and the viewers alone have control over the stream in shared game control settings. We focus on an extended *Twitch Plays Pokémon* setup (*TPP++*) and our test-bed *CrowdChess* that allows reasoning about users' decision effectiveness in such settings. Through studies, we learned that having individual influence options is appreciated, and which input mediation options and further self-orchestration features are favored by a group of people in this context. Overall, the results in Chapter 5 and Chapter 6 add to **RQ3**. Finally, the main contributions of this work are summarized in Chapter 7, where we also identify and discuss opportunities for future work.

Chapter 2

Background and Related Work

This chapter provides the theoretical and practical foundations on which this thesis builds. We first give an overview on the research methods used. Then we elaborate on the *Self-Determination Theory* and focus in particular on the benefits when users receive more choices and autonomy. Supporting this thesis' questions, it has positive effects on people's motivations and behaviors. We then present work and open questions in relation to increased users' influence in the context of games and gamification, crowdsourcing and live-streaming.

2.1 Research Methodology

In this section, we introduce the main methods we have used in the context of this thesis. Overall, we investigated empirical evidences *quantitatively* and *qualitatively* and utilized different data gathering techniques. We extensively used *online questionnaires* and complementarily used *semi-structured interviews* where reasonable. Both methods were used to elicit user needs and requirements without using prototypes or specific systems. In addition, these methods were used during and after the prototype/system usage to learn about users' perceptions of the actual realizations. Considering the former, besides investigating questions that do not require any concrete systems, this also allowed us to investigate these aspects without introducing bias, as stated by Orji et al. [227]: "*actual implementation may create additional noise as it involves many other design decisions and the results can easily be biased by specific implementation decisions*" (p. 464). When working with system realizations in user studies, we also used *observation in laboratory* or "*in the wild*" / *field studies* to learn how participants interact with the systems. Complementarily, here, we also *collected usage data* that we quantitatively analyzed. In the following we discuss these methods in more detail:

Online questionnaires: Online questionnaires are used to reach a broad user base in an economical (time and cost-wise) manner [93, 150, 252]. One advantage is a smaller participation threshold, as people can do them whenever they have time in an environment they know (e.g., at home). They can also do them anonymously, which should reduce the social desirability bias [215] (i.e., participants respond not truthfully, but with what they think is socially more appropriate). Another advantage is that the entered data is directly available digitally and can be received without the need to have an experimenter present [150]. Considering drawbacks [150, 252], while this method can lead to participants that are more diverse than traditional samples [93] it might introduce a sample bias (e.g., all participants are Internet users). Because of the reduced control (in comparison to laboratory studies) and the anonymity, it cannot be ensured that the participants stay focused during the study, and they might fabricate their answers [150, 197]. This requires a more thorough data sanity check than other methods (see [53, 197] for an overview). We used online questionnaires for various reasons:

- To validate assumptions (e.g., in Section 3.3.1, to investigate whether the *wisdom of crowds* [289] idea is applicable in the waste recycling context).
- To evaluate concepts and ideas before implementing them (e.g., in Section 4.2.2, to investigate the idea of self-tailored gamification).
- To assess users' needs and expectations (e.g., in Section 5.2.2, to receive insights into viewers' preferences towards live-streaming features).
- To complement our online-based experiments (e.g., in Section 4.3.2, to assess the intrinsic motivation and perception of game elements used).
- To assess experiences with our prototypes (e.g., in Section 6.3.2, to assess the experience viewers had while playing *CrowdChess*).

Our online questionnaires were mainly distributed through student mailing lists (consisting of psychology, computer science, media informatics and art and design students) and social media channels, and advertised through notices around campus. They were usually not incentivized. The only exception was the questionnaire we used in Section 4.5, which was published via *Amazon Mechanical Turk*, a platform on which crowd workers receive small payments for conducting tasks. This helps to reach a high number of answers quickly at a good quality, apparently independent of the money spent [30, 189]. The method sections of the corresponding studies in the next chapters will provide insights on the type of open and closed questions we asked in these questionnaires.

Semi-structured interviews: Online questionnaires are not optimal when fully detailed answers are needed: while open questions in the form of free text questions can be included, it cannot be expected that participants will spend much time on these in general. Furthermore, questionnaires do not make it possible to elaborate on specific aspects that might arise from the answers to the questions [252]. While some answers might lack clarity and it would be

helpful to ask about these aspects further, other answers might spark follow-up questions in general. Where we deemed higher expressiveness necessary, we used interview techniques. In the presented cases in this thesis (for example in Section 6.2), we decided to do the interviews in a semi-structured form. Here, the interviewer has a set of questions he or she wants to ask (similar to the *structured interviews* that resemble online questionnaires, but in which no deviations are allowed), but can deviate from this script based on the answers participants provide (similar to *unstructured interviews*, which resemble normal conversations as no questions are fixed, and thus cannot easily be replicated) [252]. In this sense, *semi-structured interviews* combine the positive aspects of *structured* and *unstructured* interviews [252].

Data analysis: Throughout the studies, we used self-created questions to investigate the aspects that were of interest given the studies' goals. Where reasonable, we also used standardized questionnaires (e.g., the *Intrinsic Motivation Inventory* [59] was used in Section 4.4). For answers to closed scale-based questions it is an ongoing debate whether the results can be treated as ordinal or interval data¹⁴. To account for this discussion, we report not only mean values but also the median value of such answers. In addition, we tested for corresponding preconditions when using further statistical methods and report non-parametric tests where the requirements were not met. For answers to open questions we used qualitative content analysis methods. For the presented studies that involved such questions (e.g., the analysis of gamification concepts in Section 4.5) we use thematic-based or content-based analysis [118] in which we derived overarching themes that emerged through the answers.

Laboratory and "in the wild"/field user studies: While the *Trash Game* was evaluated conceptually (see Section 3.3.3), all other systems were evaluated under *laboratory* or *"in the wild"/field* conditions. *Laboratory* studies give the experimenter high control over the environment and can thus reduce the amount of confounding effects, but ecological and external validity might suffer through the artificial context [65] (but see [18] for the usefulness of this type of study). We chose this method for *TPP++* (see Section 6.2.2), but tried to mimic the actual live-streaming scenario to increase the validity. In contrast, *"in the wild"/field* studies have a higher ecological validity, as users can use the system "naturally" [65], but these give the experimenter less control. Our studies can be seen as *"in the wild"* studies with laboratory aspects. For example, in the study using our gamified task management application (see Section 4.3.2), we only required participants to use the system (at least) once a day, but they could decide how. This means users knew they were part of a study, but could use the application as they wanted. Another example is the evaluation of *CrowdChess* (see Section 6.3.2). Here, we conducted the user study using the live-streaming platform *Mixer*. Thus, other *Mixer* live-streaming consumers could join anytime during the experiment, mimicking "real situations". In all these studies, we *observed* and collected *usage data*, to understand how users interact with our created prototypes.

¹⁴See for example <https://goo.gl/E2Cnt1> (last accessed: 2018-07-07)

2.2 Supporting Users' Influence

In the next two subsections, which are based on Ryan and Deci's publications [58, 255, 256], we give a brief overview on human motivation and put an emphasis on the *Self-Determination Theory* (SDT). It provides the theoretical foundation for why it is reasonable to provide people with more influence and choices. After this, we elaborate on studies reporting positive effects when choices and options are offered. Although the SDT is formulated without a game focus (and used in different contexts), it is also of relevance for games, which we will also discuss, especially in relation to *forced play*. This section is closed by discussing users' influence at design and runtime in HCI systems.

2.2.1 Intrinsic and Extrinsic Motivation

Early research on motivation only considered a binary distinction where individuals were either not motivated (*amotivation*) or motivated. Over the years, though, it was shown that the degree (i.e., how strongly motivated someone is to do something) and also the types of motivation differ (i.e., where the motivation originates from). Considering the latter, two general types are scientifically accepted today: *intrinsic* and *extrinsic motivation*. The former drives people to do something, because the activity itself is inherently interesting or enjoyable and thus brings satisfaction. An example of intrinsically motivated behavior is reading a book on a topic out of curiosity. In contrast, extrinsic motivation is described as motivation that drives people to do something, because of an external trigger. This can be external tangible or symbolic rewards (e.g., salary or grades) but also factors such as deadlines. Considering the above example, extrinsically motivated behavior could be reading a book that is not interesting but was given as a homework assignment, and not reading it would lead to sanctions at school. Many activities that people do are not motivated intrinsically, but rather through external means. Taking away the external motivation will then likely lead to stopping the behavior. Considering the example, when there is no consequence to not reading the book, it is likely that the person will not do it.

Intrinsic motivation for a task can be measured through self-reported data of interest and enjoyment of the activity through standardized questionnaires such as the *Intrinsic Motivation Inventory* [59] or the *Gaming Motivation Scale* [155] (which can also measure extrinsic motivation). Another option is using so-called *free choice* tasks. Here, users are given a task they need to conduct and after that are free to do something, including (without stating this) continuing working on the task. If they decide to continue doing the task, it can be assumed that it intrinsically motivates them to do so (as there is no external reason to do it). The more time they spend on it, the higher the motivation. Both approaches to measure intrinsic motivation are widely used in the literature (see for example [57]). Besides questionnaires, extrinsic motivation can be measured by, for example, comparing effort expended with and without an external reward given.

2.2.2 Self-Determination Theory

While other motivational theories are used in the game context (see [248] for an overview), for this thesis, we focus on the *SDT*, as it is “one of the most established theoretical frameworks within gamification and game motivation research” ([199], p. 526). A core aspect of it is the focus on three psychological needs and the relevancy of need satisfaction for psychological health and well-being. According to the *SDT*, settings that support/satisfy the individual’s need for *autonomy*, *competence* and *relatedness* foster high-quality forms of motivation.

Autonomy is the need to engage in an activity under one’s own volition and having the feeling of being in control to do activities in line with one’s own goals. Deciding on doing an activity and/or having choices in it should support satisfaction of the autonomy need and should positively impact well-being. *Competence* is the need to experience mastery, being able to acquire new abilities and having an optimal challenge. Features such as dynamic adjustments of difficulties in an activity supports the need satisfaction of competence. *Relatedness* is the need to feel connected to others and be involved in a social context. The satisfaction of this need is supported, for example, through social features in systems (e.g., a chat tool) or if group activities are supported. Regarding this thesis, the autonomy aspect is most relevant for it. Although there are more needs considered in the literature, Sheldon et al. [274] underline the relevancy of the *SDT* needs. They investigated ten needs and asked their participants to recall events in their past that they would consider as satisfying. The authors found that autonomy, relatedness and competence belong to the most important needs in these recalled events. This was also stable across cultural background and the time elapsed between when the recalled event happened and the study.

The *SDT* consists of several sub-theories that consider different facets of motivation. We only briefly mention two that underline the thesis’ aspects further. All theories can be read in more detail in [58, 255, 256]. The *Cognitive Evaluation Theory* (*CET*) focuses on autonomy and competence and considers the effects of the social context on intrinsic motivation. According to the *CET*, events that change the perceived locus of causality (relating to the feeling of self-determination and autonomy) affect intrinsic motivation. When an event provokes a change in perceptions toward a more external (internal) locus, intrinsic motivation will be undermined (enhanced). In the same line of argumentation, events that support or diminish the perceived competence impact the intrinsic motivation as well. In a meta-analysis [57], it was found that tangible and symbolic rewards (i.e., external components) can undermine intrinsic motivation, found both through *free choice* tasks and self-reported measurements.

While the *CET* focuses on intrinsic motivation, the *Organismic Integration Theory* concerns internalization and integration of values and regulations and focuses on extrinsic motivation. It states that there are different forms of it based on the degree of internalization of an activity (i.e, how far regulations have been transformed from external aspects to personally endorsed ones). Thus, extrinsic

motivation needs to be seen as a continuum (*External Regulation*, *Introjected Regulation*, *Identified Regulation* and *Integrated Regulation*) based on the perceived locus of causality, i.e., the degree to which the behavior is perceived as self-determined (relating back to autonomy). The lower end of the continuum – *External Regulation* – represents the least autonomous form, as it is considered controlled and externally regulated and can undermine intrinsic motivation (see above). *Introjected* and *Identified Regulation* differ in the amount of internalization of the regulations and in how far a person identifies with the values of a behavior. On the upper end of the continuum, *Integrated Regulation* is the most autonomous form, which shares many qualities with intrinsic motivation and appears when regulations are identified as personally important and have been assimilated to the self. The more autonomously extrinsic motivation is perceived by individuals, the more positive effects were reported (e.g., greater engagement [220]).

Overall, these sub-theories highlight the importance of self-determination and autonomy. As described, providing more autonomy has positive effects on intrinsic and extrinsic motivation. The game-related areas that we consider can be improved in terms of autonomy, by empowering users and giving them more choices to impact the experiences that unfold, as illustrated in Section 1.3. Work such as [257] discusses the relevancy of these aspects in relation to self-determination: minimization of choices is perceived as controlling and, in contrast, having meaningful choices allowing people to find something that they can endorse facilitates self-determination. In the next sections, we elaborate on studies in which it was shown that having choices led to positive outcomes, further underlining the relevancy of the thesis' questions.

2.2.3 The Effects of Choice

In this section, we present work showing that positive effects occur when people are given choices or options to exemplify the autonomy aspect of the *SDT*.

Langer and Rodin [157] considered two groups in nursing homes. In one group, it was communicated that everyone in the group can be responsible for him- or herself and still has influence on their own lives. The other group received a communication with the same content, but here, the staff's responsibility for them was highlighted. With questionnaires and behavioral measures they found differences in the groups. For example, participants of the first group reported being happier and more active, and significantly more of them attended an optional movie night. Corah and Boffa [50] exposed participants to white noise and ensured that it was perceived as equally loud by every participant. During the experiment, participants received either an escape (i.e., they should turn off the sound by pressing a button) or no-escape trial (i.e., they should not press the button). Half of the participants were given a choice. They were told that they could decide not to press the button in escape trials (when the sound was not uncomfortable) and could press the button in no-escape trials (when the sound was too uncomfortable). Through self-reports and skin response measurement

it was found that participants having a choice had reduced levels of aversive equality of the stimulus. Lefcourt [163] provided an overview of further studies in respect to different aversive stimuli and showed similar effects, in humans and animals alike. Stotland and Blumenthal [284] studied the effects of choice on anxiety reduction. In their experiment, they allowed half of their participants to choose the order of ability tests they needed to solve (the actual solving was not part of the experiment), while the other half received the task to solve the tests in a fixed order. All participants were told that the order itself had no relevance for their scores in the end. They found that the subjects without a choice were more anxious. Overall, these presented studies all showed that providing people with choices has beneficial effects on perceptions and behaviors.

Amabile and Gitomer [6] conducted a study in which children were allowed to select boxes containing material for doing handicrafts. For every child that was able to select something, another child received the same boxes from the experimenter (i.e., the child had no choice to select another box). Both children received the task to create a collage with the material. All collages were rated in terms of how creative they were. As one result, it was found that the children that were able to select the boxes were significantly more creative. Zuckerman et al. [337] considered a puzzle-based task. 80 college students were paired and only one participant per pairing received a choice. This participant could decide which three (of six) puzzles he or she wanted to solve and how much time to spend on it. The paired participants received the same puzzles and were told how much time for each of them was available (both based on the selection of the other participant). After the puzzle solving, a *free choice* phase was initiated where it was measured how long participants continued to work on further puzzles. It was found that participants in the choice condition reported significantly higher feelings of control, spent more time on puzzle solving during the *free choice* phase and would more likely return to the laboratory for further puzzle solving. Overall, this shows the positive impact of choices on intrinsic motivation directly affecting the task outcomes. Throughout this chapter, we will provide further (digital) examples, in which the benefits of choices are visible as well.

Illusion of Choice

In the studies above, participants in the choice conditions had factual choices. In another branch of choice-related research, it was investigated whether giving participants an illusion of having choices also leads to positive effects. For example, Swann and Pittman [290] conducted a study with elementary school children where they were confronted with drawing tools and other toys. In one condition, children were told that they had a choice and could pick any of the games to play with, but at the same time, the drawing game was verbally highlighted (*"Well, since you're already sitting in front of it, why don't you start with the drawing game?"* (p. 1129)). The authors excluded children in their analysis that chose not to play the game (so there was no real choice from a study perspective). In the other

condition, the missing choice was highlighted (“*I used to let children choose what game to play, but I can’t do that anymore. Instead I will tell you which game to play.*” (p. 1129)) and after five minutes, a *free choice* phase started. It was found that children with the feeling of having a choice more likely chose the drawing task and spent more time with it. Another example are the findings of Langer [156]. He showed that this illusion of choice is also apparent in situations that are only based on chance, such as lotteries. In one of his reported studies participants had to draw a lottery ticket by themselves or received it from someone (i.e., they could not draw it, but saw how it was drawn). Participants that had a choice would less likely give their ticket away and would sell it only for significantly higher prices, even though the chances to win with the ticket remained the same independent of who drew the ticket. Overall, this indicates that merely the feeling of having control and choices is sufficient to invoke effects.

Choice Overload

Another stream of research considered whether having too many choices is problematic. Schwartz calls this “*Tyranny of Freedom*” ([267], p. 85). Iyengar and Lepper [126] acknowledged that in many of the choice studies, often only a small number of choices is considered. They raise attention to the fact, that if people already have knowledge on what they prefer, they might be content with rich options (e.g., the choices in a restaurant menu), but might be overwhelmed in contexts where they have fewer insights. The authors use the term “*choice overload*” (p. 996) for these situations and state that this is demotivating. In a range of studies where they ensured that participants have no pre-determined preferences, they varied the amount of options to select from (low: six, high: 24 to 30). They found different effects, such as that in a low condition, more items than in a high condition were bought later on and that the quality of essays was better when offering fewer possible topics to write about. They found that more choices were perceived as difficult, frustrating and led to less satisfaction when participants were asked about their choice afterward. They close their consideration with the aspect that choice overload is more crucial in non-trivial choice-making scenarios and that effects might change, when choices are self-generated instead of imposed as in their studies. In general, how choices should be presented (choice overload is one aspect here as well) and how people can be supported in making a choice (e.g., through tools such as recommender systems) are topics which are considered within the area of *choice architectures*. We will not further elaborate on this, but refer the reader to [133, 298] for an overview.

Taking the previous study results together, considering this thesis, it can be hypothesized that offering more influence options in systems will also provide positive results, as long as the amount of options is not overwhelming for users. Thereby it seems to be unimportant whether the offered choices are actual or only illusory ones, as a feeling of having choices is what appears to matter. In the next section we will discuss how games are beneficial from a *SDT* perspective.

2.2.4 Games and Self-Determination Theory

Work such as [238] summarizes how games satisfy the *SDT* needs, which further explains why games are successful (see Section 1.1): competence was satisfied early through arcade games, as players could experience enhanced competence feelings while playing. Autonomy was satisfied beginning with home-based games as choices were offered to the players, for example, in role-playing games. Relatedness was satisfied through, for example, multiplayer games. Thus, as (video) games are able to fulfill these needs, they spark intrinsic motivation. Ryan et al. [258] investigated video games in relation to the *SDT* and also highlight that playing such games is an intrinsically motivated activity directly related to how much the players experience need satisfaction during play. In studies, the authors found supporting evidence, as for example in a single-player game the perceived in-game autonomy and competence was associated with the game enjoyment and the desire to continue to play. Recently, Iten et al. [125] investigated choices in narrative-rich games (an example of such a game would be *The Witcher 3*; see Section 1.1). They highlight that having choices is one feature of interactivity. They analyzed game situations in which players reported to have experienced meaningful choices. Based on this, they created a narrative and manipulated the meaningfulness (low/high) and whether there was a choice or not. They found that meaningfulness had a significant effect on player experience. Additionally, they only found a significant effect between having choices or not in the high meaningfulness condition, underlining the importance of meaning in this context as well, i.e., that not every choice matters.

Forced Play

Although games can satisfy the *SDT* needs, they do not do this in all cases. Deterding [63] investigated how the situational context in “leisurely” and “non-leisurely” game play impacts autonomy. He conducted interviews with people that engage in both (e.g., professional eSports players or game journalists). They should report their experiences of perceived low and high autonomy, choice and consequence in their interaction with games. Deterding found that in a leisure context the questions *when to play and for how long*, *what to play* and *how to play* are important factors for autonomy that are mostly not available at work. For example, a game journalist that needs to keep a deadline to deliver a report on a specific game (and thus needs to play it) cannot be considered as acting autonomously. But even in a leisure context in which multiplayer games are played as a team, autonomy is restricted when a player needs to be present. This indicates that the context matters and that forced play does not necessarily lead to the same positive outcomes as voluntary game play. Furthermore, playing at work (even though part of it) was considered as inappropriate by Deterding’s participants. Examples that are given were that one would fear not noticing a colleague having a question or could not let one’s emotions run freely, in comparison to playing at home.

Heeter et al. [110] considered the impact of forced game play in the serious games context (see Section 1.2) and stated that often “one-size-fits-all” solutions are used in a teaching context (i.e., the same game needs to be played by everyone) where pupils are forced to play the educational games. In a study with four different types of games they found that forcing game play on non-gamers or gamers that do not like the game genre/the game leads to negative effects. They add that people are highly idiosyncratic when it comes to games and that even creating a good game is no guarantor that everyone will perceive it positively. Mollick and Rothbard [203] present an overview on games at work, showing that this topic already has a long history, and focus on the issue of “*mandatory fun*” (p. 4), i.e., that games here are manager-imposed and are often aiming at supporting the employer’s goals. “*Mandatory fun*”, as discussed in Section 1.3, is not a voluntary aspect and also removes the spontaneity connected to fun (i.e., it feels like part of the work, instead of play). Gamification is seen similarly by these authors as it is imposed from the top. They highlight that consent is important to reduce the effects of “*mandatory fun*”. One aspect towards consent is an active engagement, i.e., employees should be able to decide whether they want to play or not.

These examples show that while games can satisfy *SDT* needs, there are negatively influencing factors such as “one-size-fits-all” solutions and forced play. Overall, the role of autonomy was also highlighted in these studies. This directly relates to the topic of this thesis, i.e., giving users more influence options in game-based systems. In the next section, we complement this by briefly discussing which options users receive at the design time and runtime of systems in HCI, showing that the idea of empowering users is reasonable in general.

2.2.5 Users as Designers

Sanders [261] highlights that people like to feel creative and want to be creators. She discusses different levels of creativity and that people should be empowered to express this. Along these lines, Fischer and Scharff [76] stated that “*designing a system that can sufficiently anticipate all possible uses in advance... is an impossible task*” (p. 396) and that “*providing the opportunity for people to become designers is both important and rewarding*” (p. 398). They stress that not everyone wants to be a designer, that even if users are open to this, there are different degrees of engagement to consider and that there needs to be a distinction made between design time and runtime of systems. The latter is relevant, because here, users can discover mismatches between what they want and what the system offers.

At Design Time

User-centered design [252], *co-design/co-creation* [262] and *participatory design* [151, 266] are methods that consider later (end-)users’ opinions and needs at design time. According to Sanders and Stappers [262], in *user-centered design* users are seen as subjects. Here, they are often not directly involved in the design tasks,

but studied from an ethnographic perspective to learn about their needs, requirements and behaviors. In the other approaches users are seen as partners [262] and directly involved in the design process [88]. Nonetheless, in all these design approaches users have the chance to influence the later systems, although the designers or researchers can, depending on the chosen approach to a varying degree, moderate what will be integrated eventually [85]. Vines et al. [312] provide an overview on the goals that motivate user participation in the design processes in HCI. These are sharing control with users, sharing expertise and motivating individual, organizational and technological change. Based on their argumentation, the first aspect relates to the autonomy aspect by giving groups of people the chance to impact the later systems which they will need to use. The second aspect relates to the fact that users in many domains of HCI understand their current practices and potential issues better than external researchers. Thus, giving them the chance to articulate these helps researchers to elicit knowledge, values and opinions that can be added to the system's design. The third aspect helps to motivate changes already through the design process. The authors state that users should receive more options than they have today, for example, by being integrated in how the process and tools to be used should look. Thus, the authors call for more influence options, fitting the thesis' idea to empower users, although we focus on user influence at runtime.

Several studies show positive results of integrating users in the design process. Kanstrup [135] provided an overview on different forms of integrating the user in the design process: *end user programming*, *lead user innovation* and *participatory design*. While they have similarities, there are differences in, for example, the motivation for why end users are integrated in the process. Overall, though, in all these cases, users are enabled to design their own systems. They further report findings from design workshops conducted with 17 families who were asked to design IT services supporting everyday living with diabetes. Here, participants reported a broad range of design ideas that revealed that end users are in general able to derive ideas and are competent designers, even without having an IT background. Gerling et al. [88] compared the performance of wheelchair users and game design experts in designing wheelchair-based games. They found that both groups were able to create concepts providing valuable insights into the design of such games and elaborated on important design aspects (such as mechanics or aesthetics). The authors found that non-experts miss details (e.g., technical constraints or missing relationships between game mechanics), unlike experts. Gaye and Tanaka [85] described a case study in which young people conceptualized and developed an interactive information pack in a *Do-it-Yourself* approach. While researchers here had only a guiding function, the young people designed and implemented the system themselves. The resulting prototypes were positively received by them. One reason for this was the strong ownership the participants reported. Taken together, these aspects show that users are already having an influence on the design time of systems and can also successfully impact the design with positive outcomes.

At Runtime

Considering giving users influence on systems at runtime, Sundar [285] argues that the most seductive aspect of modern communication technologies is their potential for customization, i.e., the degree of user autonomy and/or how many options the user receives to modify or create content. Kiouisis [144] considers several definitions of interactivity, and many of these highlight the choices that are available to users in a system. He states that “*individuals should be able to manipulate the content, form and pace of a mediated environment in some way*” (p. 368). Marathe and Sundar [185] put emphasis on the aspect that customization places the locus of control within the users. They mention several aspects as to why customization is used by users, amongst them to adapt a system towards their own goals, making it more efficient for them, or helping in managing complexity or information overload. Systems can offer various degrees of customization and through this, users can also more likely start to identify with a product. The authors conducted a study where users could customize a news platform in terms of which content was shown, as well as the appearance of the content and the web page. They found that participants in the customization condition provided higher values for sense of control and sense of identity compared to participants who had no chance to customize, underlining the effects customization produces. This directly connects to the thesis’ idea of giving users more influence options at the runtime of gamified systems and game live-streams.

In later works, Sundar et al. [287, 288] highlight that interactive tools can serve as a source for others and offer a form of self-expression that goes beyond customization. They call this “*source interactivity*” ([288], p. 2248), which relates to autonomy and control. They see this as a continuum. While graphic or functional customization is part of it, on its higher end they see that users are even able to create content on their own. Examples they provide are *Wikipedia* and *YouTube*. Here, users would not only make a conscious choice of what they read or see, but are also actively creating content as a source. Fischer [75] considers *cultures of participation*. He highlights that users should be supported in being consumers and designers. System designers should make users *co-designers* by providing elements that allow these users to adapt the system to their needs during runtime. The *culture of participation* is supported by making changes seem to be possible and beneficial, and to have low barriers for users. He also highlights that not every user wants to be a designer and thus, they should have the opportunity to contribute only if they want to. Considering the introduction on games and play (see Section 1.1), customization in game-related approaches seems beneficial as well. Campbell et al. [34] stated: “*narratives, goals, and challenges can benefit from design that leaves a maximal number of choices open to the player. By placing fewer restrictions on interaction, play can occur at the pace of the player. This means rules should be designed to shape play without being overly restrictive on when and where play can take place or how the player chooses to explore the narrative of the game*” (p. 250). All this highlights that users should be able to choose how to interact with systems, directly adding to the idea this thesis follows.

2.2.6 Summary

In this section, we introduced the concepts of intrinsic and extrinsic motivation and the *Self-Determination Theory*. Based on these elaborations, our research questions receive theoretical grounding: empowering users with more options in systems should positively impact extrinsic and intrinsic motivation according to this theory, foremost as the feeling of autonomy should be higher. We reported on studies that showed positive effects of having choices and discussed how games, when not forced, in general facilitate intrinsic motivation. We also shed light on practices that empower users at design and runtime, showing that it is beneficial to give users more influence options in general and that this is already something that is considered in HCI. With this thesis, we add to this, as we provide users with more autonomy at the runtime of gamification settings and game live-streams. Given the previous sections, this seems reasonable to do, and we hypothesize that this impacts the motivation positively and thus the users experience. To our knowledge, this has not been considered to the extent this thesis does. In the next sections, we highlight the current research state and open questions in this regard. In Section 2.3, we elaborate on why it is useful to tailor gamification and that personalization and customization are two approaches for this. While customization approaches exist already, they do not give users full control over the gamification in a system. This is something we consider in Chapter 4. In Section 2.4, we give an overview on crowdsourcing, and elaborate on how games and gamification are an option to motivate participation. We discuss how the actual crowd members profit off the systems they put work into and highlight that this can be improved. We consider crowd-based systems, in which the user effort directly improves the system for the users in Chapter 3. In Section 2.5 we consider the live-streaming context and discuss motivations for why people watch these and what is done to empower viewers' interactive options. Here, it becomes obvious that the interactive design space is not fully understood yet. Our contributions to this are presented in Chapters 5 and 6.

2.3 Tailoring Gamification

In their literature review, Seaborn and Fels [269] highlight that gamification is seen as a tool that is able to facilitate extrinsic and intrinsic motivation. Study results reported in the literature were mixed or positive. The mixed results were attributed to, among other things, differences in the perception of gamification. In this section, we complement the introduction on gamification given in Section 1.2.1 by focusing on the efforts towards tailoring it.

The relevance of context factors in the perception of games was mentioned previously (see Section 2.2.4). The impact of individual factors was also reported on in the game literature. For example, the influence of personality on the liking of particular game genres [231] or which aspects in online gaming are relevant

for someone [129] were reported aspects towards individual differences. Thus, it is unsurprising that “one-size-fits-all” approaches in game-related sub-fields also have shortcomings [25]. Many works stress the negative aspects of such approaches, calling for tailored solutions (e.g., [145, 223]). Several aspects were found to have an impact on the perception of gamification elements and game-related aspects such as personality traits [45, 131, 196], player types [204, 306], age [22, 137, 228], gender [45, 148, 228], achievement goal orientation [101], social factors [105], culture [4] and the application domain [106].

As we consider personality and player types in the study presented in Section 4.5, we will elaborate on these further. Different models exist to measure the personality (e.g., the *Myers-Briggs Type Indicator (MBTI)* [210]), but the *Five Factor Model (Big Five)* [91] “is currently the leading and most widely [used] model adopted by personality psychologists” ([227], p. 460). It considers five main personality traits on which people can score low or high. One way to measure these is through questionnaires, such as the *NEO-PI-R* [51] or the *Big Five Inventory* [242]. The literature [225, 242, 306] describes the traits in this way: high scores on *Openness to experiences* represent people that have the tendency to be curious and creative and have an adventure- and knowledge-seeking attitude. People scoring low tend to be conservative, tend to have fixed views and do not want to gain new experiences. People high in *Conscientiousness* tend to be self-disciplined, goal-oriented, organized and give activities thought before doing them. Low scores represent people that tend to be careless and unsteady. People high in *Extraversion* have the tendency to be outgoing, seek out new opportunities and are social. People low on this trait tend to be introverted, silent and withdrawn. People high in *Agreeableness* tend to be altruistic, considerate, tolerant and caring. People low on this trait tend to be critical and distrustful. Finally, *Neuroticism* describes emotional stability. People high in this trait tend to be nervous, fearful and depressed. Low-scoring people have a higher emotional stability.

Ferro et al. [73] discuss relationships between personality and player types. They state that these just have different contexts (e.g., personality has a general one, while player types are specific to games). They relate player type models, personality aspects and game elements/mechanics and show that there is overlap between the different concepts. Hamari and Tuunanen [107] conducted a literature review in relation to player types. They found that different segmentation methods have been used so far. One is psychographic segmentation in which, for example, users are only distinguished between casual and hardcore players. Another segmentation is on a behavioral basis, i.e., how users behave within a system. An example for this is Bartle’s work [11], in which player behavior in a specific game genre was observed and four player types (*Killer*, *Achiever*, *Socialiser* and *Explorer*) were identified. Hamari and Tuunanen examine the approaches and found that many concepts considered in the literature were overlapping (e.g., the concept of achievement was found consistently). Various other player models are currently considered in the literature (see [31, 73, 227, 303] for overviews), such as the *Demographic Game Design Model* [12] or the *BrainHex* [212].

One recent development in this respect is the *Gamification User Types Hexad Scale* [306] developed by Tondello et al., which is based on the *Hexad framework* by Marczewski [186]. In contrast to other available player type questionnaires which aim at games in particular, it focuses on systems using gamification [306]. The *Hexad framework* is based on the SDT needs and contains six user types which are associated with people's intrinsic and extrinsic motivation: the user types *Socialiser* and *Philanthropist* have similar characteristics (relating both to the SDT need for relatedness). While the *Socialiser* wants to interact with others and create social connection (a pure relatedness aspect), the *Philanthropist* wants to help others and is thus characterized as being motivated by purpose. The *Free Spirit* and *Disruptor* are motivated by the SDT need for autonomy. While the *Free Spirit* is characterized by having freedom in a system and acting without external control (a pure autonomy aspect), the *Disruptor* wants to trigger change in a system and alter it. Finally, the *Achiever* and *Player* relate to the SDT need for competence. While the *Achiever* wants to complete tasks and prove him- or herself, the *Player* is motivated by earning extrinsic rewards. Tondello et al. [306] showed that the user types derived with the *Gamification User Types Hexad Scale* relate to game elements and *Big Five* personalities. Importantly, the scale provides scores for every player type. Thus, it counteracts the criticism of other models which often treat users as belonging to just one type [107].

Works which consider the relationship between player types/personality traits and game elements approach this topic differently, either by investigating a particular element in depth or by considering a broader spectrum. A recent example for the former is Jia et al. [130]. They investigated different leaderboard constellations and found, amongst other results, that people scoring high in *Extraversion* experience leaderboards more positively. A recent example for the latter is Tondello et al. [303] who investigated the relationship between the *Hexad* user types and game elements. They let participants rate a set of game elements in terms of their enjoyment through an online study. Based on this, they created eight element clusters and relate these to player types, personality traits, age and gender. Through this, they were able to show differences, such as that the cluster *Customization* is preferred by women and men who scored high on *Openness*, or that the cluster *Risk/Reward* was preferred by younger *Achievers* and *Players*.

Considering these examples, it becomes obvious that much is being done to understand factors that influence the perception of gamification. A question is how these theoretical findings can be practically applied to systems that want to provide a tailored experience. Theoretical results gained can be used through at least two options. The first option is personalization. Here, either algorithms automatically derive the relevant factors and adjust the system for the users, or the system is adapted to a target population by developers beforehand [31]. The second option is customization. Here, users can adapt the system themselves. The findings might be used in a recommender system fashion [305, 304] to assist users here. Personalization and customization will be considered in the next sections, but given the thesis scope, we only briefly discuss current personalization efforts.

2.3.1 Personalization

Monterrat et al. [205] present an architecture to realize a personalization approach in a gamified learning environment. Here, individual game elements would automatically be activated or disabled based on user data (e.g., player type and age), usage data (e.g., interaction within the system and interactions with the game elements) and environmental data (e.g., the context where the system is used). Elements could also be disabled by the user, being a customization aspect in an otherwise personalized system. As the authors state, this is not only done to allow for more freedom in the system, but also to learn more about user preferences in particular. In a later work, Monterrat et al. [204] showed positive effects of a personalized approach based on player types in a learning context. Six experts considered five offered gamification elements for the system (e.g., the game uses a set of stars if a player learns a new grammar rule), and mapped them to player types. Other participants were now confronted with either the two best matching gamification features in the learning platform (given their player type) or with the worst matching. They found that participants with matching features spent 39% more time on the platform, showing positive effects of personalization. Böckle et al. [24] propose a design framework to create personalized gamification approaches. Which adaptivity criteria to consider is the most significant aspect here and covers, amongst others, player type/personality type, how users use the system, who the users are and the context. This further underlines the number of aspects that play a role in tailoring gamification.

Orji et al. [224] used storyboards in the context of risky alcohol drinking behavior depicting different (game-related) persuasive strategies (amongst them, goal setting, competition, cooperation, personalization, and customization), which “are the 10 commonly employed strategies in persuasive games and gamified system design” (p. 1016). The authors were able to show that personality traits have an impact on the perception of these strategies. They propose a “one-size-fits-all” approach to appeal to a broad range of users, by using aspects that are motivational for some personalities and not negatively associated to others. Additionally, based on their results, they discuss how to tailor to specific personality types, such as that people high in *Conscientiousness* should be offered (among others) goal-setting elements. In a recent study [226], they did a follow-up with the *Hexad* user types in the same context and with the same strategies. They showed that the player types have a significant impact on the perception of the strategies. Again, they derived a “one-size-fits-all” approach appealing to a broad user base, as well as ways to tailor it for particular user types. This also highlights that there is not yet a perfect approach that appeals to all player types similarly. In [222], Orji et al. showed how their results can be applied in a tailored serious games context and that tailoring indeed had positive effects. Jia et al. [131] investigated several game elements (amongst them, points, leaderboards, clear goals and challenges) and their relation to the *Big Five*. They showed videos of an interactive prototype (in the context of promoting healthy habits) which implemented the game elements. Similar to the earlier approaches, they were able to derive

strategies on how to use game elements to appeal to a broad user population (e.g., using levels) and how to tailor the system for a specific personality type (e.g., people high in *Extraversion* more likely prefer points, levels and leaderboards). Overall, these aspects have only considered particular aspects in tailoring. As mentioned previously, the literature reports on more aspects that have an impact on the perception of gamified scenarios (besides player types and personality traits). To our knowledge, though, there is no approach that considers all these aspects for a personalized system yet.

Khoshkangini et al. [142] present an approach in which challenges are created through procedural content generation in a personalized fashion, based on the player's preferences, history and former performance. In the context of a smart urban mobility game, these were appreciated by users. Busch et al. [31] considered different *BrainHex* player types in a game setting and created tailored game missions for these. They hypothesized that the user experience should be affected by this (e.g., the congruent missions should lead to better player experience), but could not show it. This also indicates that more factors need to be considered. As an issue, they stated that "*there are no empirically-driven design guidelines on how to translate player types into meaningful game designs*" (p. 160). Similar aspects were raised by Böckle et al. [25] in a literature review conducted on personalized gamification. They identified several research challenges, such as that is necessary to learn how the individual reacts to different (game) mechanics, how to balance the degree of adaptivity and how to design meaningful adaptive gamified systems. Overall, these aspects show that there are still a range of open questions regarding how to personalize gamification interventions.

2.3.2 Customization

Following the argumentation above, the range of challenges in personalization and the aspects that impact the perception of game elements makes personalization currently difficult. A different approach to achieve tailoring is customization, i.e., allowing users to adapt the system. Considering the *SDT*, this directly connects to the autonomy and choice aspects discussed previously (see Section 2.2).

Personalization vs. Customization

Orji et al. [225] considered personalization and customization. They found that both aspects are perceived positively, although users perceived personalization in their study context (healthy eating and risky alcohol drinking) better. Through a qualitative analysis of participants' answers they were able to derive strengths and weaknesses that both approaches share (e.g., they increase the relevance and usefulness of a system), but also differences. Customization gives users a feeling of control, freedom and the ability to add a personal touch to the system, but they also found that participants see issues: it is expected to divert attention, to consume more time, and is assessed as difficult. Furthermore, participants seem

not to trust their own judgments to decide what is best for them (in the health context here). This calls for improved user interfaces and assistance tools to mitigate these effects. One example is recommendations based on derived player types that can then be further customized by the participants [304]. Considering the drawbacks of personalization, they found that participants judged it as potentially boring and also highlighted trust aspects. In a recent work, Orji et al. [226] found that customization is not negatively associated to any of the *Hexad* user types, while personalization is negatively associated with *Disruptors*, showing further differences.

Although not in a game context, Sundar and Marathe's [286] comparison of personalization and customization add to these aspects as well. They discuss the question that in a tailored context, it is unclear whether users are more satisfied because they have more autonomy and feel as "*self-as-source*" (p. 299) (hinting that customization might be more valuable) or because the content is now a better fit (hinting that personalization alone would be sufficient). They conducted a study in a news page context. In the customization condition, participants received the option to customize which menu items are available on a news page (a comparatively simple form of tailoring); in the personalization condition, the page was adapted by researchers for the participants, based on the participant's preferences (which were collected earlier), and no tailoring was a baseline condition. They were able to show that power users appreciated customization, while non-power users seemed to appreciate personalization more. In a follow-up, they also considered privacy-related aspects based on the idea that personalization can only be done when the system collects data in the background. Here, they found that power users favor customization in scenarios with low privacy (i.e., where it was highlighted that the system uses browsing information) and personalization in high privacy settings (i.e., ensuring the participant that the system will not collect any data). Overall, these approaches show that customization and personalization have both advantages and disadvantages and factors that influence their perception, making it reasonable to investigate both approaches. Within this thesis, we contribute to the customization literature.

Meaningful Gamification

Another important aspect towards customization in gamification was highlighted by Nicholson [218, 219]. He, based on the *SDT*, states that when gamification is perceived as meaningful and relevant, it might more likely produce autonomous, internalized behaviors. He emphasizes that it is necessary to put users in the loop to achieve meaningful gamification. They should be able to decide on *what* should be gamified to reach their goals, and define these by themselves. They should be able to decide *how* the gamification should look (e.g., which game elements are used) and *why* they should use a gamified system (e.g., they should be able to relate game elements to the underlying activities). To achieve this, Nicholson demands that users need to be involved in the creation of gamified systems

(i.e., highlighting a strong *user-centered design* approach) and/or by allowing customization of the system itself. He suggests to create systems in which users can “create their own tools to track different aspects of the non-game activity, to create their own leveling systems and achievements, to develop their own game-based methods of engaging with the activity and to be able to share that content with other users” ([219], p. 4). In a later work, Nicholson [220] further discussed that users should find their own reasons for engaging with a behavior or system to support the intrinsic motivation. He introduces six concepts to achieve meaningfulness. Two of these, which he called *Play* and *Choice*, are directly connected to the ideas of this thesis. With *Play*, the voluntary nature and freedom of choice is highlighted. He states that players should be able to “establish and change their own constraints”, that “play must be optional” and that “players need to be able to select what they want to play with” (all p. 5). *Choice* emphasizes putting the player in control of how he or she engages with the system. This directly relates to the autonomy aspect of the *SDT*, as the player always should have the choice of which activities he or she wants to do and how to engage with a gamified system.

Overall, this directly connects to the aspects we discussed in Section 2.2 and shows that it is also advisable to consider users’ opinions at design and runtime and to give them the chance to alter the gamification.

Forms of Customization

We give examples of game-related customization in this section, which illustrates three aspects. First, the positive effects of choices (see Section 2.2.3) are also visible here. Second, customization is used in game-related contexts already. Third, different degrees of customization in systems exist.

Customization on the element level: The following examples show contexts in which users were able to customize individual elements: Kim et al. [143] differentiate *functional customization* (i.e., users can alter something that has an impact on the game outcome) and *aesthetic customization* (i.e., users can alter the audio-visual experience). In their first study, they allowed for functional customization (i.e., participants in this condition could make their ship in a 2D shooter game stronger). In their second study, they allowed for aesthetic customization (i.e., participants could alter visual features of a car in a racing game). Compared to not being able to adjust something, they overall showed that players in the customization conditions had stronger feelings of autonomy and control leading to higher ratings for game enjoyment. Peng et al. [233] investigated how more autonomy and competence-related features in a game scenario are perceived. For autonomy, they provided users with the option to customize the in-game avatar’s gender and appearance (aesthetic customization), character attributes (e.g., causing more damage; functional customization) and allowed users to talk with non-player characters by offering several answer options in the dialogs (instead of having no options). For the competence aspect, they offered a dynamic difficulty mechanism, give feedback on how well a

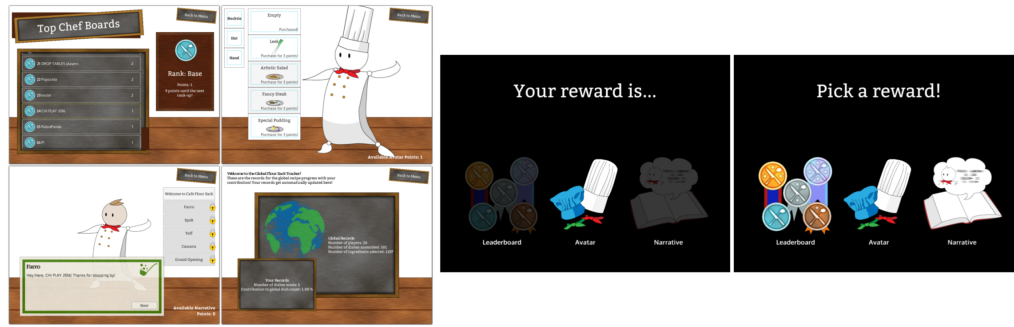


Figure 2.1: Parts of the user interface of *Cafe Flour Sack*. Left: The different reward-related screens. Right: The system selects a reward randomly, or a user can select a reward (all images taken from [276]).

player performs in the game and provide achievements. They could show that participants exposed to the customization options had greater satisfaction of the autonomy need, enjoyment and motivation for future play. The same was found for the competence features.

Mollick and Rothbard [203] conducted a study in which they created a game with two themes (fantasy or farming). Depending on the condition, participants were either able to select the themes (preferred choice), were assigned one of the options (no choice) or were asked what they liked but received the other theme (non-preferred). Overall, the authors could show that participants that had a choice scored higher on several self-reported metrics, e.g., they understood the rules better, found the game to be more fair, reported to be more engaged in the game, and that the offered choice directly predicts the feelings of autonomy. Munson and Consolvo [207] considered the elements, goal setting, rewards, self-monitoring and sharing in a physical activity context. They found that goal-setting (a form of customization) was especially important, when users could set the goals for themselves. The gamified task management applications considered in [136] allowed users to add tasks and to assign points to these. When the task was finished and checked off in the application, the points were unlocked and, for example, would improve a virtual character. Here, users can thus decide on their own how “worthy” their tasks are. Birk et al. [21] allowed users to customize their avatar, leading to higher degrees of identification and need satisfaction. Recently, Birk and Mandryk [23] showed that it also has positive effects on attrition in self-improvement programs. In [272] users could set up their own competitions in a physical activity context, which led to an increase in daily steps.

Customization on the game configuration level: Another customization approach is to allow users to adjust which game elements they want to use. Siu and Riedl [276] considered different reward schemes (see Figure 2.1, left) in a categorization game (called *Cafe Flour Sack*): leaderboards (users were ranked based on their points they could achieve for correct categorizations), customizable avatars (receiving a currency that can be spent on digital items to customize a 2D avatar), unlockable narratives (receiving a currency that can be spent to unlock

short stories) and a global progress tracker (showing a player's contributions in relation to all contributions). They conducted a study in which some participants received a reward randomly and others could decide which reward type they wanted to receive (see Figure 2.1, right). They found that the choice in this system led to better task completion times and similarly-engaged player experiences. While the users were not able to further customize the actual rewards (i.e., the amount of points received after a correct categorization), this work underlines the positive effects when users can decide which game element they want to use.

Snow et al. [280] investigated a gamified tutoring system in which users could receive points that could be spent to unlock game-based features (e.g., the capability to edit an avatar). This can be seen as another example in which users can decide how game elements are introduced. In this context, the authors found negative effects on learning (e.g., as it diverts students' attention). In a study by Schöbel et al. [265], participants were confronted with ten different elements for a learning environment context. They saw several subsets and combinations and ranked the elements. Overall, level, points, goals and status scored highest. In addition, they stated how many elements of the ten they would combine in the learning platform. The answers showed that there is a high variance, making it reasonable to allow users to customize this aspect as well.

Customization on both levels: To our knowledge, gamification systems that offer a broad range of customization options that combine both levels extensively have not yet been investigated. Here, users should be able to decide on all aspects at runtime, even to the degree of disabling all game elements completely. They should have a large set of game elements to select from, and be able to combine these and further customize each of them. In this sense, users could decide in a "bottom-up" way on their gamification setup. We see this as self-tailored gamification, which fits into Nicholson's [218, 219] considerations for meaningful gamification (see above). This thesis contributes studies of such approaches and shows positive qualitative and quantitative effects of them in Chapter 4.

Users' Openness to Customization

Given the criticism that customization might be too demanding for users, the question arises how it can be supported, when they have even more choices in gamification approaches where they can decide on many aspects in a "bottom-up" way. At this point, it needs to be highlighted that there are areas in which users are already keen on spending considerable effort in a "bottom-up" manner. One example is the *Quantified-Self* movement [43, 175]. Here, users decide themselves on which data they want to monitor (e.g., daily food intake), how they want to do this and what they do with the information. Choe et al. [43] showed that there are different reasons for why people engage in this movement, such as the desire to improve health or gain new life experiences. They stated that a form of *self-experimentation* is involved and although people might lack scientific rigor here, HCI tools can support people in this task.

The figure shows a three-step wizard for creating a game.
Step 1: General Information (1/3) includes:
 - **Name***: A text field with "Doctor in the House" entered.
 - **Description***: A text area with a placeholder text about being a doctor.
 - **Thumbnail**: A small image placeholder and an "Upload Thumbnail" button.
 - **Playing Instructions**: A text area with a placeholder text about the game's theme.
Step 2: Topics and Questions (2/3) includes:
 - **Topics***: A list of topics (Acne, Anemia, Mumps) with an "Update Topics (csv)" link.
 - **Questions***: A list of questions for each topic (e.g., "What are the symptoms of [topic]?") with an "Update Questions (csv)" link.
Step 3: Advanced Settings (3/3) includes:
 - **Order**: Radio buttons for "Random", "Fixed by question", and "Fixed by topic".
 - **Access**: Radio buttons for "Public" and "Private".
 - **Reward**: Radio buttons for "Encourage original answers" and "Encourage popular answers".

Figure 2.2: The game creation wizard in the *Games for Crowds* platform (taken from [98]).

Lee et al. [162] studied the concept of *self-experimentation* in the context of sleep education control. Here, they evaluated the idea of using not only interactive instruction materials, but also a just-in-time intervention tool. The former included a protocol in which participants define a behavior that they want to achieve; set a goal they want to achieve; generate ideas for how to reach it with behavior change techniques (with material that helps them to select and personalize techniques); formulate a final plan; and use self-tracking measures to determine if the goal was reached. The just-in-time intervention tool uses off-the-shelf hard- and software and allowed users to set up events based on a simple rule-based approach that should trigger them at certain points in time where this is beneficial for reaching their goal (e.g., a reminder at 9pm that informs them to go to bed). In their study, they found improvements in sleep quality for both aspects. Considering self-tailored gamification, we also see the relevance of *self-experimentation* as users can experiment with the game elements as well.

The literature also showed that people are open to creating game-based experiences in general. Guy et al. [98] developed a platform (called *Games for Crowds*) that allows employees to create their own games. They can voluntarily decide to create games through a wizard component (see Figure 2.2) to reduce complexity, which also targets the criticism above. They can define general information (such as a game description), topics and questions, the order for which questions should be presented, who has access to the game (all employees or only a subset) and the kind of rewards (e.g., points for completing a task, or bonus points if an answer was given by several other players already). In a three-month “in the wild” study, 25 employees created 34 games, of which 24 were business-related and ten were leisure games. Overall, 339 players played with the games. They found that there were different motivations for why games were created, such as that creators wanted to gain or expand knowledge, discover assets or ideas, teach the crowd or produce fun and engagement. The game creation itself was assessed as easy overall. Only the transformation from the problem-to-be-solved into a game was assessed as difficult. Interestingly, more choices for creating the games were demanded. The games generated useful data for the company, showing that the creators were able to develop successful games on their own.

Again in a working context, Mollick and Rothbard [203] showed that employees have a tendency to self-create games here, i.e., employees start to use game elements on their own, to make their activities more rewarding. One such example is provided by Donald Roy [253], who observed machine operators doing repetitive and monotonous work. He reported that they decided to voluntarily add simple games and game elements to their work (for example to set up a personal goal, such as doing a certain work activity many times in a row) to make their tasks more fun, without receiving any further reward for this. In a non-working context, Massung et al. [190] investigated a monetary gamified version, a gamified and a baseline version of an app that aimed at encouraging people to log whether shops have an open or closed door (in relation to energy savings). They also report that many participants in interviews stated that they used a form of “self-gamification” strategies, such as setting goals/challenges themselves (e.g., doing more work in the app than the day before). This behavior was not designed for in the first place nor anticipated a priori. A similar aspect was recently reported by Steinberger et al. [282]. They conducted a study in the context of how to create gamified applications that make safe driving more engaging. They found that “*many drivers come up with their own gameful experiences*” (p. 2835). All the examples in this section show that users are willing to spend effort and are open to create their own game-based settings.

Similarly related are “modding communities” of games. Scacchi [263] states that this is a form of user-led innovations and a *Do-it-Yourself* approach to customize and tailor games. As Postigo [237] highlights, those individuals that create modifications for existing games (with the help of the developers through tools) spent considerable time on the programming or creation of assets. Postigo investigated motivations for those people, and among them is the motivation to be able to identify with the game and the desire to create their own experience. Also, scientific approaches such as *PingPong++* [331] provide users means to modify/create games on their own, even without being experienced programmers.

2.3.3 Summary

In this section, we elaborated on tailoring gamification as “one-size-fits-all” gamification solutions often provide only mixed results, as there are a number of individual and contextual differences. While personalization and customization are both options to achieve a tailored experience, we focused on the latter, as it fits the overall thesis question: through enabling users to alter the gamification at runtime to be a better fit, they can exert influence and should be able to act more autonomously. While we presented customization approaches in gamification, to our knowledge, a fundamental “bottom-up” gamification approach (where users could decide on game configurations and game elements) has not yet been investigated so far. This thesis contributes to this area by investigating such approaches. The findings we obtained in this respect are reported in Chapter 4. Additionally, in this section, we showed areas in which “bottom-up” considerations are carried

out by users despite the increased effort (in comparison to personalization or using “one-size-fits-all” tools). We also presented related approaches in which users started to establish game or gamification settings on their own as well. These examples show that people are willing to spend time and effort to establish game-based experiences, even if not asked or required to do so. Thus, allowing users to create game-based scenarios is not far-fetched and has been reported in different cases already. Consequently, allowing and supporting users in creating their own gamification approaches seems reasonably grounded.

2.4 Crowdsourcing

Crowdsourcing (also known as *collective intelligence*, *user-powered systems* or *human computation* [66]) follows the idea of the *wisdom of crowds* [289], i.e., that a group of people comes to a better decision or more reliable result than an individual. Crowdsourcing is used for many different aspects that are not (yet) easy to solve by computer systems: because the problems are not solvable with current hardware, it would not be efficient, or humans outperform computers in the corresponding task [56, 122, 206]. Additionally, crowdsourcing is also used to enhance Artificial Intelligence algorithms [38]. Doan et al. [66] stated that “*it appears that in principle any non-trivial problem can benefit from crowdsourcing* (p. 87). Geiger and Schader [87] categorized four different forms of crowdsourcing approaches:

- **Crowd solving:** The heterogeneity in the crowd is used to generate a number of different solutions to a problem that has no pre-definable solution, such as ideation or design contests (e.g., as can be found on *DesignCrowd*¹⁵).
- **Crowd creation:** A large number of individual contributions create a comprehensive experience (e.g., user-generated content). Heterogeneity of the individual contributions is important. Examples for *crowd creation* are *Wikipedia* or *YouTube*.
- **Crowd rating:** Approaches use the *wisdom of crowds* idea [289] in particular to generate high-quality results based on homogenous decisions of the crowd. Here, a group of people receives the same task and if they provide the same solutions, this can be seen as a quality metric. An example of such a task is image tagging (which is elaborated on below).
- **Crowd processing:** Tasks are divided into chunks that can be solved by the crowd (not all tasks need to be solved by the same person). An example would be a task on *Amazon Mechanical Turk* in which, for example, 400 company names needs to be completed with their address data, to be searched for on the Internet. In this task, obviously, while one person can solve this alone, by distributing it to a crowd of people, the tasks can be solved faster.

¹⁵<https://www.designcrowd.com> (last accessed: 2018-07-07)

As this thesis has a gaming and not a crowdsourcing focus, we only briefly elaborate on aspects that are essential for the game-based systems in the next chapter, namely motivation in crowdsourcing, image-based tasks and aggregation methods. For an overview on further crowdsourcing aspects, we refer the reader to the overview article of Doan et al. [66], Lasecki et al. [158] and Malone et al. [183] and to literature reviews on the usage of crowdsourcing, such as [87] and [206]. In these sources, more crowd-based systems are presented and challenges in this area are discussed (e.g., how to recruit contributors, how to distribute tasks, or context maintenance questions).

2.4.1 Motivation in Crowdsourcing

How to motivate people to participate in crowdsourcing tasks is an important topic [66]. Typically, users can decide whether they want to participate in a specific crowdsourcing task or not [87]. On platforms such as *Amazon Mechanical Turk*, users have a range of tasks they can select from. Following the arguments made in this chapter, this is beneficial for motivation. Geiger and Schader [87] considered how users can be matched with tasks that are suitable for them. They highlight that this is beneficial for motivation and the result quality. Another aspect for motivation is task length. Lasecki et al. [158] highlight the usage of microtasks in this respect, i.e., small, context-free tasks that can be solved in a few seconds. These were shown to be beneficial for motivation as well, but not all problems can be formulated as such a task [296, 308]. Microtasks, although having longer overall task completion times, lead to fewer mistakes, greater stability in respect to interruptions, and are easier for the users [40]. They are also often used in crowdsourcing [87], for example in collaborative writing [295] or image-labeling task: as *ExpenseControl* and the *Trash Game* both contain image-based microtasks (see Chapter 3), we present related approaches below. Doan et al. [66] mentioned several examples for further encouragement and retention schemes in crowdsourcing, such as instant feedback for users, creating an enjoyable experience, allowing competitions among crowd members and creating ownership situations in which users feel that they own parts of the system. Partially, these are game-based aspects we will elaborate on next.

Payment and Game-Based Approaches

Morschheuser et al. [206] conducted a literature review on crowdsourcing and also highlight that there many reasons for why people participate in such tasks, either extrinsically or intrinsically motivated. Considering extrinsic motivation, an example is monetary compensation [206]. Mason and Watts [189] investigated paid tasks on *Amazon Mechanical Turk* and found that a higher payment increases the quantity of work done, but had no effects on the quality, whereas the kind of compensation scheme (e.g., pay per work step or pay per task completion) has an effect (i.e., pay per task completion leads to more effort expend and better

quality). Other work provided evidence that money might also have a negative influence. For example, in a crowd-based application in which information could be shared on whether or not shops have a closed door, paid users (in comparison to users that were not paid to participate) were less likely to participate further if no money would be spent anymore [190].

Tasks that are assessed as promising for the development of one's own skills are considered intrinsically motivating [189]. Also, game-based approaches are often used in crowdsourcing to spark intrinsic motivation. Examples of this are the *ESP game* [315] (an image tagging game; see below) or *Verbosity* [316] (a *Taboo*-like game to collect common-sense facts). The studies presented in the respective papers showed that these games attracted users to solve many microtasks. Thus, these games can be considered as intrinsically motivating (see Section 2.2.1). Further support for this comes from a study conducted by Eickhoff et al. [68]. They compared a paid non-game with a paid game version of the same categorization tasks. They found that through the game-based approach many crowd members spent more time and solved more microtasks (which had no further effect on the money they could receive), less cheating occurred and the game was replayed more often. Morschheuser et al. [206] also highlight the relevancy for gamification in crowdsourcing, although they emphasized that most systems only use simpler gamification and that it would be beneficial to employ more sophisticated ones. In a literature review conducted by Seaborn and Fels [269], the relevancy of gamification as a motivator in crowdsourcing was also underlined. Feyisetan et al. [74] investigated gamification in paid crowdsourced image labeling tasks and found that gamification has further positive effects, as the quality and the amount of added tags improved considerably. Kobayashi et al. [147] analyzed different motivational elements in an unpaid crowdsourcing task, among them gamification, which also showed positive effects. This shows that crowdsourcing is a context in which it appears reasonable to use and investigate game-based aspects. Considering this thesis, for our crowd-based systems (see Chapter 3), we thus decided to use such approaches.

Benefits for the Crowd

To illustrate another motivational aspect, as an example we present a crowd-based system fitting the category of *crowd solving* above: in *Chorus* [161], users can ask questions (e.g., *What should I buy my nephew?*) and a crowd in the backend starts to generate ideas and discuss options that can be upvoted by the crowd members. For the asking user, only one result is presented and it appears that he or she is talking with just one person, instead of a crowd (see Figure 2.3). *Chorus* uses a point- and payment-based reward system (mapping to financial compensations) in which bonuses are granted when an answer is selected that a crowd member has upvoted or has suggested him- or herself. Although crowd members can be inspired by such discussions to pursue similar aspects (in the given example, buying the same object for a relative), in general, they have no

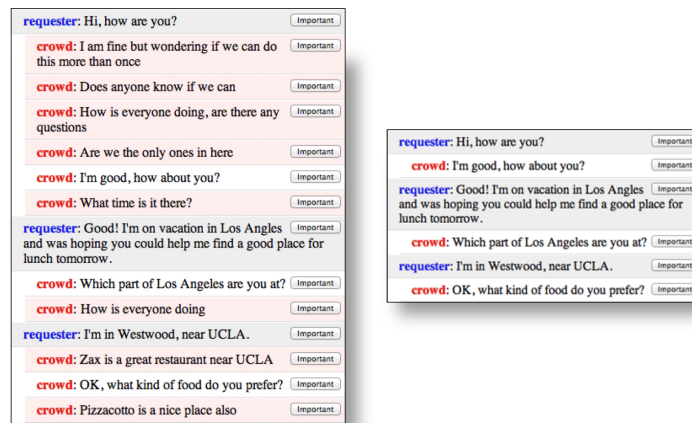


Figure 2.3: A chat example from *Chorus* (extracted from [161]). Left: Unfiltered view, messages in red did not receive enough crowd upvotes. Right: How the chat appears for a user asking a question.

direct benefit from their contributions besides the reward that is implemented for motivation in this system. This relates back to the meaningfulness aspects discussed in this chapter. In this example here, the task solving is of no further consequence for individual members, besides whether or not they receive their monetary compensation. Interestingly, work discussing motivational options for crowds (for example, Malone et al. [183]), mentions money, enjoyment, socializing with others, the knowledge of contributing to a cause larger than oneself and receiving recognition, but does not mention meaningfulness or reciprocity (i.e., receiving something back from the system besides money) in particular.

In the next section, we provide examples for *crowd rating* systems that have a similar setup with crowd members not directly profiting off their contributions themselves. Considering *crowd processing*, certain tasks could be imagined that provide something to the crowd members. For example, tasks at work might be distributed to several colleagues to solve them faster. *Crowd creation*, on the other hand, with the given examples, appears more meaningful inherently and more rewarding for the crowd contributors. For example, on a knowledge platform adding information improves the underlying system through the crowd. Nonetheless, contributors do not directly benefit from their own contributions, as they know what they have done and would not receive more knowledge by re-reading their own text on *Wikipedia*, for example. In contrast, their contributions might spark interest in others to become contributors themselves who, potentially, would add aspects that are also of interest for the original contributor. This is a common reciprocity scheme in social networks [42, 105]. The initial contributor, though, only profits from his or her contribution indirectly. In a study, Wasko and Faraj [321] found that there are also other factors that motivate people to expend effort in these contexts (e.g., improving their professional reputation). Nonetheless, it was also shown that here the *Pareto principle* applies as well [147], i.e., often, 80% of the effort is done by only 20% of the people.



Figure 2.4: VizWiz and ESP game. Left: Pictures, questions and answers in VizWiz (extracted from [20]). Right: ESP game interface (taken from [315]).

We consider game-based, crowdsourcing systems (see Chapter 3) in which crowd tasks are directly interwoven with the service the system provides. We will not use external crowds. Instead, the system's users can solve the tasks and thus are directly profit from their own contributions. This should add meaning to the actual tasks they need to solve. As discussed above, the game aspects and the added meaning to the tasks should positively impact the user motivation. Similar to the *crowd creation* systems, we see these as self-sustaining systems because the system service is only reasonably possible through user effort. The direct coupling between one's own effort and the benefits one receives from the systems adds a particular aspect to it. Through the significant user influence on the system's outcome, this fits with the overall question of this thesis.

2.4.2 Image-Based Tasks

In this section, we present approaches that used image-based microtasks. These are relevant and related to our self-sustaining systems in the next chapter.

Image Labeling

Image-based microtasks are often used in systems in which the answer of the crowd is necessary during runtime of the actual system [158]. One prominent example is *VizWiz* [20] which is an assistance tool to support blind people. A user can take a picture and can ask a question (e.g., "What bottle is this?"). A paid crowd sees the taken image, interprets it and answers the question. The answers are read back to the user in the system. Figure 2.4 (left) shows example questions and crowd answers. *VizWiz* performed well and is an alternative to expensive (and specialized) assistance hardware [20]. Another example is *PlateMate*, a crowd-based nutritional analysis tool. Here, a paid crowd receives pictures of user's plates, before and after a meal. Through a series of tasks (locating food items in the image with bounding boxes, identifying the food in every box textually, and

estimating a portion size) the system generates (with the help of a food database) the meal's calories, fat, carbohydrates and protein and was shown to be nearly as accurate as a trained dietitian. Both examples show that image-based microtasks solved by a crowd have potential in general.

As a method for motivating users to provide labels for images without monetary compensation, von Ahn and Dabbish [315] developed the so-called *ESP game*, a game with a purpose (see Section 1.2). Here, a player sees a picture and needs to enter a descriptive word for it (see Figure 2.4, right). Another player also inputs tags for the same picture and only if both players manage to enter a matching tag, they receive points. Thus, the system receives image tags while the game is played. The system uses matching words from previous rounds as taboo words. When the same image is used again, taboo words are forbidden to use, ensuring that the same words are not entered repeatedly through the system. The *ESP game* implements several other aspects, such as anti-cheating mechanisms and spelling checks to ensure high tag quality. In a four-month period where the game was accessible on the Internet, nearly 300,000 images were tagged with more than 1.2 million labels of good quality. 80% of their players played on multiple dates, showing the intrinsic appeal of this game and that a crowd of people is able to generate meaningful data playfully. In a later work, von Ahn et al. [317] presented another game-based approach, *Peekaboom*, which aimed at recognizing objects in images. Initially, one player only sees a black image. The other player sees the complete image (e.g., an elephant) and a word (e.g., trunk) and needs to arrange for the other player to guess this word. He or she can reveal parts of the image and can give specific hints (e.g., "A noun"), but no further communication is possible. While the first aspect helps to find objects in the image, the second aspect classifies these further. The game uses points, a leaderboard and achievements. In a month's runtime more than 14,000 people played the game and generated more than a million pieces of data, with an overall good accuracy. All these examples show that a crowd can be used to classify images and can be motivated by game-based approaches. These are aspects we built upon in *ExpenseControl* and the *Trash Game* in the next chapter.

Text Recognition

Optical character recognition (OCR) has improved significantly in recent years [216] and is used in various domains, for example to digitize printed documents [115], to translate text in real time [307] or as assistance technologies for blind and visually impaired users [28]. Nonetheless, it is not yet error-free [20]. OCR results heavily depend on the quality of both the picture taken and the printed text in the picture [335], especially in contexts where the picture is taken with a mobile device [70], because of bad lighting and/or picture distortions. Further issues occur, for example, with pale ink [318], old typefaces, bad scans or crumpled paper [44, 335]. In the literature, approaches are presented that use crowd-based systems to enhance the outcome of OCR.



Figure 2.5: Examples of OCR-related tool. Left: Example captcha of *reCAPTCHA* (taken from [318]). Right: *Digitalkoot* interfaces (taken from [44]).

With *reCAPTCHA*, von Ahn et al. [318] presented an extension of the *CAPTCHA*¹⁶ system that is used to distinguish humans from computers, commonly used on web pages. While the latter uses computer-generated randomly rendered images, *reCAPTCHA* uses OCR results (e.g., old issues of the New York Times from 1860). In both cases a user needs to read and decipher the shown word(s) to proceed. This is a simple microtask that was assumed to not be solvable by computers. Von Ahn et al. consider image parts of scans with two different OCR programs. When the outcomes do not match it is presented together with a control word for which the answer is already known as a *CAPTCHA* (see Figure 2.5, left). The correct entering of the control word allows the user to proceed. The answers to the other image part are compared across multiple users to gain a higher confidence level, to generate new control words and to use it as a correction for the OCR result. In tests, it was found that two to six users would already be sufficient to decipher a broad range of images correctly. They also found that the performance of *reCAPTCHA* was better than the performance of the OCR programs and could compete with professional transcribers, without adding more effort than the original *CAPTCHA* system. After one year, 440 million words were deciphered, showing the power of microtasks in the OCR context.

Game-based motivation for crowds to do OCR corrections were also investigated. *Digitalkoot* [44], for example, used archive material from Finnish Newspapers from the late 19th century in two games: in the first game, participants see images and the corresponding OCR results. For every pair, a player is asked to indicate whether or not these match (Figure 2.5, right). In the second game, players have a time limit and need to build a bridge to save mole from falling down and dying by typing the word they see in the image (Figure 2.5, right). If the typing was incorrect, the newly built bridge part would explode. In both games players receive points for correct answers and lose points for incorrect ones. The system uses verification tasks (i.e., tasks for which the answer is known) to identify malicious players and to reduce the system latency of giving feedback. These tasks are created automatically when seven players agree on a result in the game. The authors considered 51 days of the system’s runtime and in this time frame 2,740 hours were played and 2.5 million tasks solved. Overall, they checked

¹⁶ Completely Automated Public Turing test to tell Computers and Humans Apart.

the accuracy of *Digitalkoot* in a random subsample and it was nearly perfect (in two sample articles with 1,467/516 words, only 14/1 error(s) were made, while OCR alone produces 228/118). Similar to the approaches above, it shows that crowdsourcing for these kinds of tasks is possible. Similar to the *ESP game*, though, users had no benefit to themselves from playing the game, besides the entertainment value of it. In the next chapter, with *ExpenseControl*, we present a system that directly integrates OCR-related microtasks in its design, and doing these tasks improves the service the users receive from it.

2.4.3 Aggregation Methods

In areas in which a group consensus needs to be found, the question is how to mediate different opinions. While such aspects are investigated in group decision-making in general (e.g. [120]), they are also investigated in crowdsourcing in particular. Here individual contributions need to be combined to come to a result [66, 158], especially in the presence of potentially contradicting opinions [158]. For example, in a *crowd rating* scenario in which five crowd members classify an image as X and one other crowd member as Y, there needs to be a decision on the validity of X and Y. In this dissertation, aggregation of inputs is also a relevant topic. First, they are considered in the next chapter to maximize the outcome of the self-sustaining systems. Second, they are of relevance for the game live-stream scenarios in Chapters 5 and 6, as here also individual contributions need to be mediated. Hung et al. [239] highlight that not only the different expertise levels of humans, but also the difficulty level of the task-to-be-solved are both issues for why contradicting answers in crowdsourcing can appear. The authors differentiate non-iterative (i.e., interactions such as answers are treated separately) and iterative aggregation technique (i.e., interactions to previous instances will be considered for future instances as well). They created a test framework to evaluate several aggregation techniques in terms of computation time, accuracy, robustness to spammers and adaptivity to multi-labeling (i.e., whether these can also be used for more than binary decisions). They found that depending on the goals a system designer pursues with the crowdsourcing approach, different aggregation techniques might be more relevant. For example, when a fast answer is required, an aggregation technique that works on a plurality voting scheme (i.e., take the answer that is provided by most of the users) is suitable. If time is of no consequence, an aggregator based on expertise values of the users works best, i.e., users that performed well previously should receive higher values in, for example, weighted plurality votes.

Lasecki et al. [160] developed *Legion*, a framework that allows end users to capture existing user interfaces that are live-streamed. A crowd is able to issue mouse or keyboard inputs to operate it in real time. Input aggregators were implemented and tested in two different scenarios: controlling a robot (offering only a few different commands the crowd can select from) and a spreadsheet transcription (offering a large input space). Different aggregation approaches were considered:

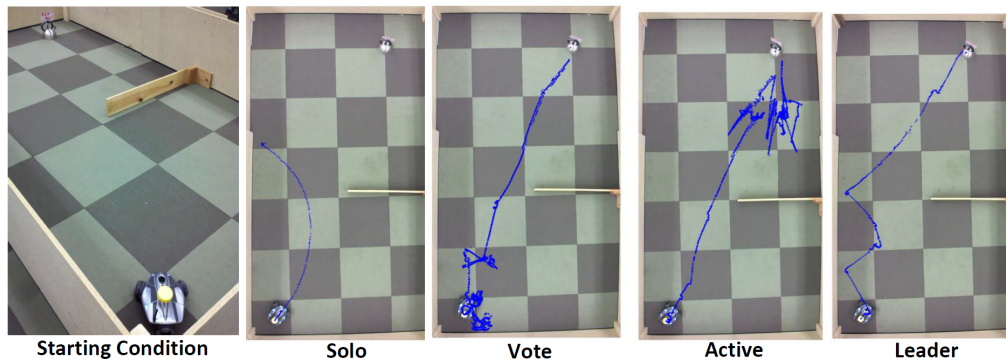


Figure 2.6: Examples of robot movements in the navigation task in the *Legion* study with the different input aggregations (taken from [160]).

one aggregator (called Mob) serialized all commands and carried them out in the interface and another aggregator (called Vote) represented a weighted majority vote with weights based on the agreement of the crowd (i.e., how similar one user's input in a time frame is to all other user inputs) and only the action with the best score in a time frame was carried out. But individuals also received sole control: one aggregator (called Leader) selects the crowd member with the highest agreement value and provides him or her with the sole control, as long as this crowd member remains the one with the highest agreement value; another aggregator (called Active) provided sole control randomly as long as the crowd member continues to provide inputs at all and finally, an aggregator (called Solo) selects one crowd member to have sole control, but when this member stops providing inputs, no other member is selected.

Every aggregator was utilized in ten trials in the corresponding scenarios (see Figure 2.6 for an example of the robot controlling scenario). The results revealed an overall good success rate. Task completion time was lower for Leader than Active. Furthermore, aggregators performed differently depending on the actual task context. Overall, this work highlights that even in crowd settings, several general options on how to aggregate input exist: all contributions are considered (e.g., Mob), individual contributions are aggregated first (e.g., Vote) or individuals receive full control (e.g., Active). They differ in terms of "trust" and "latency". The first aspect covers whether many opinions or only one opinion is used. Considering the *wisdom of crowds* [289] idea, the group opinions should perform better. The latter aspect covers the issue that for an aggregation a system needs to wait for group opinion, while an individual can directly interact.

Salisbury et al. [260] considered whether a crowd is able to operate an unmanned aerial vehicle. They also considered four aggregator methods which were compared on different scenarios, in terms of accuracy, reaction time and computational complexity. They used Mob and Leader (see above); a weighted majority vote operating on time-slices, with weights based on the conformity to the other crowd members (called Real-Time Majority, which has similarities to Vote above) and an aggregator in which all available actions are ranked over all crowd work-

ers (called Real-Time Borda). Different attributes were derived: Mob was assessed as unreliable and Leader was too slow in terms of reaction rates (as it was based on only one person). While Real-Time Borda performs faster, Real-Time Majority was deemed as a better choice when accuracy is important in a task. This work shows again that the aggregation method needs to be evaluated in respect to the task to be solved.

Overall, these approaches show that aggregation methods have an impact on the outcomes of crowd-based systems and they need to be selected with respect to the goals. While we aggregate individual inputs in Chapter 5 to reduce information overload for the streamer by using a plurality voting scheme, we specifically consider different aggregation mechanisms in Chapters 3 and 6. In Chapter 3, we evaluated these to maximize the accuracy in the *Trash Game* and *ExpenseControl*. In contrast, in Chapter 6, we allowed the crowd to decide which aggregation mode should be active at runtime. This was something that was not possible in the scenarios above; here the aggregation mode was fixed and could not be changed by the crowd itself.

2.4.4 Summary

In this section, we have briefly introduced crowdsourcing and highlighted aspects that are relevant for this thesis, i.e., motivation, image-based microtasks and aggregation methods. The first two aspects are particularly relevant for Chapter 3 where we consider the effect of self-sustaining systems. Here, the users of the system solve image-based microtasks knowing that this directly impacts the usefulness of the system, i.e., the users have a fundamental influence on the system's outcome. This is a difference from the typical crowd-based approaches in which the crowd members usually do not profit from their own contributions. While we also consider aggregation methods there, these were more relevant in the shared game control settings (see Chapter 6). Here, we allowed the group to switch the aggregation mode as they see fit, to learn which benefits such a raised influence level provides. With this and the self-sustaining system idea, we can investigate how increased user influence on the underlying system is perceived.

2.5 Live-Streaming

In this section we complement the introduction on live-streams given in Section 1.2.2. We present work showing that interactive and social aspects are driving factors here. Nonetheless, these works also highlight that the live-streaming platforms need to improve to further facilitate these aspects. Following this, we then present approaches that investigate improved interaction options. While these consider the usual case with a streamer present, we also elaborate on shared game control settings, also in the live-streaming context without streamers. Here, as viewers need to self-administrate, interactivity needs to be supported as well.

Interactivity and Social Components in Live-Streams

Tang et al. [292] investigated the mobile streaming apps *Meerkat* and *Periscope*. Streamers here build a personal brand which is supported by the interactions with their audiences. Many of the activities found in the streams were interactive in nature, such as chatting with viewers or doing an Q&A session. Haimson and Tang [100] considered live-streamed events on *Facebook Live*, *Periscope* and *Snapchat*. They found that immersion, immediacy (both aiming at providing the notion of viewers “being there”), interaction and sociality (with the streamer and viewers) are dimensions that make remote event viewing engaging. Interactivity was seen as a key component. It occurs through chat messages to which the streamer or other viewers react, but also through the option to influence what the streamer shows (e.g., by asking him or her to focus on other areas of the event). The volume and content of the messages were seen as challenges that can lead to frustration. They conclude that live-streaming leads to active spectatorship that should be further supported by the platforms, for example by grouping viewers based on shared interests. Lottridge et al. [180] investigated mobile live-streaming behaviors and motivations of teens. Considering the apps that were used for live-streams, it became apparent that those were often connected to social networks (such as *Facebook Live*). They see this as a core aspect underlining that live-streaming is becoming more social and personal. They also conclude that live-streaming has changed from broadcasting-only to being interactive.

Lu et al. [181] consider live-streaming in Asian regions. While differences from Western regions are reported, the social and interactive aspects are similar. Wohn et al. [329] investigated why viewers donate to streamers, a particular form of streamer-audience interaction. They report on different motivations (e.g., paying for entertainment when they enjoyed what the streamer does or helping the streamer to improve his content) and found that a parasocial relationship (i.e., viewers build a relationship and emotional closeness with an actor although they have never met personally) is correlated with emotional, instrumental and financial social support. They considered that interactivity (which might have a further effect on donations) in live-streams already happens when a streamer verbally reacts to viewers and makes eye contact with the camera. They also conclude that the current platforms are not suitable for direct streamer-viewer interactions and need improvement, especially when channels become large.

Interactivity and Social Components in Game Live-Streams

Game live-streams are seen as a source for entertainment [95, 279]. Hilvert-Bruce et al. framed it as “*live-streaming began as a niche, gaming oriented domain, but is diversifying and growing into a broader social media trend*” ([112], p. 59). The experience of viewers of gaming-related content, be it co-located or distributed, has a history, based on the consideration that “*there is a strong sociability to gaming... [meaning] that watching gaming is a key component of play*” ([297], p. 1559). Downs

et al. [67] also highlight the social aspects of gaming and that besides the players, non-active participants in the form of an audience should also be considered. In their work on co-located play, Tekin and Reeves [297] reported that spectating is more than just watching someone play. Spectators seek to display continuous engagement with the player, up to the point where they coach the player; they criticize play techniques, recognizing and complimenting competence and reflecting on past play. Thus, the authors see a difference in “being a spectator” and “doing spectating”. Hu et al. [121], in the context of Chinese streaming platforms (with one also having a gaming focus), showed that viewers in the live-streaming context identify with the streamer. They found that this is motivational for continuous watching and that audience participatory options help to enhance this relationship. They conclude that current live-streaming platforms can improve to support the level of identification between audience and broadcaster. They suggest, for example, that streamers should be enabled to show their gratitude by providing viewers with special badges and to provide more roles to viewers to enhance the influence between groups. Pellicone and Ahn [232] investigated what makes streamers successful. They also highlight the need for future streaming platforms that provide easier ways of building communities as they assess the current platforms as not optimal in this respect.

Sjöblom and Hamari [277] used an online questionnaire to investigate why people watch others play on *Twitch*. They highlight that watching play leads to less autonomy in comparison to playing games by oneself, but at the same time has a social component to it that is not available in single-player contexts. One of the main results of their study is that these social factors are highly important, as the sense of community relates to how much people watch and how many viewers follow and subscribe to the streamer. They conclude that not only do the games need to be more appealing for spectators, but also the platforms, as a chat is not enough for many viewers. This directly connects to the question this thesis will investigate, i.e., how can we improve the current platforms and which interactive options viewers want. In a later work, Sjöblom et al. [278] investigated the relationships among video game genres (e.g., action or sandbox games), stream types (e.g., doing a *Let's Play* or a *Speedrun*) and viewer gratification on *Twitch*. The authors found that the type of stream is more important than the game played. They further found individual and contextual differences, underlining that “one-size-fits-all” interaction patterns might not be reasonable in live-streams either and a broader range of options should be offered. Further support for this aspect comes from Cheung and Huang [41]. They consider online sources on how people talk about their experiences with a streamed real-time strategy eSports game. They found a broad range of reasons why people are interested in watching. They identified nine personas (see also Section 5.2.2), and what entertains these types of spectators. Although the authors highlight that viewers can have multiple aspects of these personas, interactivity is not always a necessity. This hints that interactivity should be an optional aspect in which viewers can decide to engage.

The live-stream context is relevant for this thesis, as a formerly non-interactive experience (watching a stream) had become more interactive by integrating viewers through different means, i.e., users received influence options. While different live-streaming contexts exist, as shown, given the overall thesis question, we focus on game live-streams. The works presented here showed that the streamer-audience interaction channels can improve and it even seems reasonable to do this, given that the content (i.e., games) that is watched by viewers is also (originally) interactive in nature. As shown in the previous works, live-streaming is also a social experience that also profits from the enhanced interactivity. Based on these social aspects, this is a group setting and it becomes a relevant question how to offer interactive options in general. This thesis contributes to this by investigating interactive options on a larger scale.

2.5.1 Empowering Audience Interactions

Understanding audience interactions helps to shape performance and enriches the experience [244]. Ways to empower an audience (i.e., adding interactive options) have been investigated in different contexts so far. We briefly elaborate on some of these, as approaches and findings are also used in live-streaming.

Television

Social TV, as mentioned in Section 1.2.2, had the goal to make television more interactive and social [37] and can be seen as a kind of predecessor of interactive live-streams. An example of the former is to give viewers the chance to influence the narrative of a show. Johns et al. [132] for example, investigated a *BBC* show in which viewers were offered choices in an interactive episode featuring a small group of British soldiers. The authors highlight that this was only pseudo-interactive, as every choice was a previously prepared decision and was checked against what the episode's author deemed as correct. In a user study, they found that viewers still had a higher sense of autonomy and reflected more. Vorderer et al. [319] investigated a similar case. Participants were to watch a movie and depending on the condition they were in, they had the option to impact the story of a movie (by selecting from multiple choices) three times, once or not at all. These story adaptations were only small ones, so that overall the movie variants were nearly the same for all participants. Their results showed that cognitive capacities (indicators used by the authors were whether a participant had graduated from high school, and participants' response times) moderate how choices are perceived: participants deemed to have lesser capabilities (having not graduated and exhibiting longer response times) evaluated the movie more positively with no interactivity, while those with greater capabilities appreciated the interactivity and reported higher entertainment. The authors reason that those viewers felt more involved with the story. Overall, besides showing that interactivity can have positive effects, this study also showed viewer differences. Ursu et al. [310]

also investigated how TV storytelling can become more interactive and found elements for viewer participation that are also transferred to web videos such as voting in contests, content suggestions between episodes and the option to evaluate user-generated content for developing a story. As shown in Section 5.2.1, these are aspects that are also visible in game live-streams today, showing that these transfer to other contexts as well.

Other work considered communication tools for the audience to enhance social aspects which are also partially available in live-streaming. Geerts [86] compared voice and text chat during TV consumption. While voice communication was more natural and made it easier to follow the program, the text chat was preferred by younger viewers. Weisz et al. [323] also investigated text chat in live-streams of (normal) movies. They could show that while the chat was perceived as distracting for some participants, it also added to the engagement. This held to the degree that for some participants watching was more enjoyable, especially when the movie content was suboptimal. They also found evidence that the chat as a tool impacts participants' liking of and closeness to other participants.

Performances and Events

Empowering audiences in performances or events has also been investigated. For example, Cerratto-Pargman et al. [36] considered audience participation in a theater context. The theater play was not only shown to a local audience, but also streamed. During the performance, the audience received questions at the end of scenes, in relation to these. Audience members could answer, for example, via *Twitter* or directly on the streaming page. In addition, the audience was able to send messages during the play, which were shown on a large display on the stage. Overall, visitors had mixed feelings about an audience being integrated into the performance (e.g., because of the distracting component), although the actual impact of the audience was limited (e.g., the actress did not consider the messages during the performance). The authors conclude that "*degrees of participation in the context of interactive performances is still a rather unexplored concept*" (p. 616). Friederichs-Büttner et al. [81] allowed a co-located audience to directly interact with the play, making them co-authors. Through interviews it was highlighted that interactions should be interesting, easy and should add to the experience in a valuable fashion. Also here, the disruptive effect on the play was reported. Reichl et al. [246] provided several prototypes providing features in the context of opera live-streaming which could also be used by co-located audiences. For example, when switching rooms, a location-based application provided a summary on what happened in the opera. They also empowered the audience to change which camera perspective should be live-streamed, which was appreciated by the audience in this context.

In a movie theater, Maynes-Aminzade et al. [192] empowered a physical audience to play different games together: they allowed participants to play *Pong* by instructing the audience to lean left or right in their seats to move the paddles,

and in another set of games, the audience was able to use laser pointers to mark areas on the cinema screen. While the *Pong* example can be seen as a sort of shared game control, as the audience members shared the control of the paddle, the other games are similar to “normal” multiplayer games where every audience member had some influence. The authors concluded with guidelines, such as that the control mechanisms for audience integration should be easy, that cooperation within the audience should be facilitated and that it would be acceptable if only a subset of the audience could be sensed, as long as the audience thinks they have an influence (connecting back to the illusion of choice; see Section 2.2.3). Curmi et al. [52] investigated whether spectators can motivate athletes during running. They created a system in which online spectators receive live data of a runner such as the runner’s heartbeat, meters covered, and the runner’s location on the map. Spectators were able to motivate the runner by remotely cheering and it was found that spectators cheer at points where the athletes seemed to need it. The runner perceived these cheers through vibrations and a speech synthesizer verbalized the name of the cheering spectator. Additionally, spectators could post comments that could be inspected later on. The authors found that runners liked this kind of feedback and found it motivational, independent of whether it came from friends or unknown spectators.

Game Live-Streams

Considering game live-streams, different options for audience interactions are investigated in the literature: using platform-offered features for interactivity; showing particular inclusive behavior in the stream, potentially by also using streamer augmentations; and allowing the audience to alter the streamed game.

Platform-offered features: Live-streaming platforms provide features for audience integration. The work in this thesis was mainly done before these features were offered. In August 2017, *Twitch* launched their extensions concept¹⁷. With it, it became possible for streamers to customize their channels and allow viewers to directly interact with interactive elements below the video stream or as an overlay on the stream. One example is an extension in which viewers can hover on specific parts of the streamed video and receive information about the underlying game’s state¹⁸. This shows that today, in theory, a tight coupling between streaming page, games and viewers is possible. From the conceptual view, though, it is not clear what such extensions should offer, as, to our knowledge, so far no study was done that investigated which range of features is attractive for viewers. This is something we consider in Chapter 5. In addition, from a technical viewpoint, *Twitch* (and also *YouTube*) has the issue that there is a delay (“lag”) between streamed content and what viewers see. It is at least 12 seconds (usually more) and varies between viewers [333]. Chat messages (and also inter-

¹⁷ Engadget: *Twitch streamers will soon customize their page with new tools*, <https://goo.gl/pZ5Duq> (last accessed: 2018-07-07)

¹⁸ For an overview on this extension, see <https://goo.gl/21HXW3> (last accessed: 2018-07-07)

actions with extensions) have nearly no delay. This leads to the situation that, for example, when a viewer writes something in the chat, the streamer can directly react, but verbal responses will only be seen later in the stream for the viewers. This makes interactions problematic¹⁹. The platform *Mixer* was created to focus on streamer-audience interaction²⁰. *Mixer* was acquired by *Microsoft* in August 2016, gained traction since then and released an updated version of their API in 2017. *Mixer*, in contrast to *Twitch*, offers lag-free streaming already. Furthermore, streamers can also customize their channel's page with input elements. At the time of this writing, these options are not yet as powerful as the extensions on *Twitch*. Overall, all this shows that the commercial platforms see potential in empowering audiences and provide means that now can readily be used. At the same time, it seems necessary to explore the design space scientifically, to learn how interactivity should be realized. This thesis adds to this in Chapter 5 by investigating viewer perceptions of a range of different live-streaming-related elements and by investigating improved interaction channels.

Although the platforms provide further options today, the primary interaction channel in streams is still the chat (see Section 1.3). Hamilton et al. [108] interviewed streamers and many of them reported that communication with an audience of ≥ 150 viewers is hard to maintain. This makes it questionable how good the chat is for audience integration. Olejniczak [221] investigated *Twitch* chat messages of different sized channels and found significant differences (such as the message length or emoticon usage). He stated that *"the 1000 [viewer] sample was characterized by very long message uptime, which encouraged the chat users to interact with each other and participate in meaningful exchange of opinions and thoughts"* (p. 332), which was not the case for even larger channels. While Hamilton et al. [108] compared the situation here to a stadium, Musabirov et al. [208, 209] compared it to a sports bar where viewers could switch between roaring and talking. Ford et al. [79] call this *"crowdspeak"*: *"crowdspeak may appear chaotic, meaningless, or cryptic. However, we discovered 'practices of coherence' that make massive chats legible, meaningful, and compelling to participants. By coherence, we simply mean that the chat makes sense to participants and is not experienced as a breakdown, overload, or other difficulty"* (p. 859). Overall, based on this, the chat seems to provide certain means for interactivity, even in larger channels.

Scientific approaches consider how to improve the communication situation. *TwitchViz* [229] aims at making large chats more manageable. It supports streamers and also game designers (that might learn important facts regarding their game being streamed and commented on by viewers) in analyzing the chat history. With it, streamers and game designers receive support for post-hoc analysis of what kind of discussion happened during the stream (e.g., amount of chat messages in relation to horror scenes). With *Rivulet*, Hamilton et al. [109] investigated how streams can be combined. Through this, viewers can watch several

¹⁹ *Twitch* is testing a new option recently to minimize the lag;

see <https://goo.gl/Bqmesk> (last accessed: 2018-07-07)

²⁰ <https://mixer.com/about/story>, (last accessed: 2018-07-07)

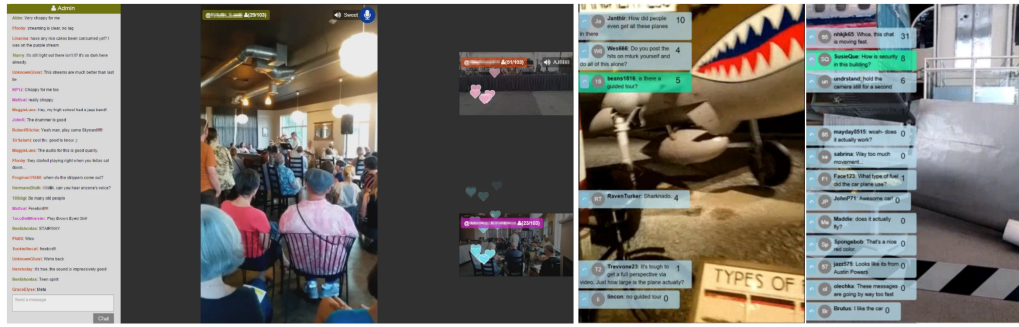


Figure 2.7: Alternative streaming tools. Left: Viewer’s view of *Rivulet* with three streams (taken from [109]). Right: Viewer’s view (with and without partitioning of messages) of the conversational chat circles (extracted from [200]).

streamers simultaneously and thus see, for example, different perspectives of the same event. Viewers can select one stream as their primary view and other streams are shown with smaller previews in parallel (see Figure 2.7, left). *Rivulet* uses a chat which combines the individual stream chats. It allows users to give “hearts” (as an easy to submit and interpret form of feedback) that are visualized on the respective stream directly, and allows push-to-talk messages that are played in the channels. At a Jazz event, the authors found that streamers interact differently with their audience and that viewers switched streams regularly. The heart feedback was used extensively and also here, information overload was reported for the chat. Push-to-talk messages were not often used by the viewers, but the content of those messages differed compared to the chat messages.

Miller et al. [200] focused on how to improve the chat in live-streaming by using conversational circles. With this, the authors aimed to solve the information overload issue. They found that simple upvoting of messages is still difficult for users when many messages are shown, which is why they aimed at reducing the amount of messages a viewer sees: viewers of a live-stream are dynamically partitioned and see only messages in their “neighborhood”. These can be upvoted and the more votes, the further the message gets distributed. This ensures that important messages become visible to all viewers. Following this, viewers and streamer are also always seeing the top three messages (see Figure 2.7, right). In a user study, it was found that such an approach is indeed feasible and reasonable, as the ease of use and the amount of messages that can be handled is increased. Furthermore, it has positive effects on the community, as less lurking occurs, i.e., people are more inclined to participate.

Overall, these approaches show that not only are the platform vendors improving features towards streamer-audience interaction, but this is also a topic for HCI research. Here, not only are novel features studied, but also existing ones are considered in terms of how these are perceived and how they can improve. This thesis adds to this by considering which options are currently used in streams keen on integrating viewers, what viewers actually find interesting on a larger scale and by also providing new interactive options (see Chapter 5).

Streamer-centered interactions: Streamers' themselves can facilitate interaction, which is another line of research. Scully-Blaker et al. [268] conducted a small-scale study in the context of *Twitch*. They found that even people that were never in the role of a streamer before adapted to the *Twitch* game streaming context as they started to interact with their audience. They state that "*participants felt as though they were playing for their audience, both in the sense that they had an imperative to be entertaining and in the sense that they felt the need to make choices that were more interesting*" (p. 2030). This underlines the role the streamer has in interactivity. Smith et al. [279] highlight interactive aspects used today in *Let's Play* streams. Viewers can have a more active part than passively spectating, which is an incentive for them: through live-chats they can suggest which game should be played next; the streamer answers their questions; and they can give hints if the streamer has missed something in the game. Additionally, a form of co-authorship happens when the streamer plays user-generated content in streamed games. Hamilton et al. [108] conducted interviews with *Twitch* streamers and viewers and also found that viewers are integrated already. They can play against the streamer in competitive games; they can provide answers through the chat that are used in a streamed quiz game; polls are used to make decisions in games, or for answering unrelated questions; and submitted fan-art is shown by the streamer in the stream. It was emphasized that this helped viewers to identify with the stream and to become regulars.

Gandolfi [83] highlights that the streamers interact with their audience in different ways, from isolated (in which a streamer only plays) to collective play (in which the spectators give tips and advice through the chat and thus become part of the show). An analysis of play sessions revealed different performances. For example, streamers in "challenge-oriented" play focused on the game play, and the ability of the streamer appeared to be the main attraction for viewers. Here, little interaction with viewers happened. Another example is "imagination-oriented" play sessions which offer a bi-directional flow between streamers and audiences, such as that both parties talk about the games played. Overall, though, the amount of interaction between streamer and audience appeared low, especially in more popular streams. Gandolfi sees an explanation for this in the "*constant and chaotic flow of messages and posts*" and in that "*these shows are seen more as top-down spectacles than interactive sessions*" (both p. 76). He further highlights that interactivity also depends on the games the streamer plays. Those games that offer a greater autonomy (e.g., the sandbox games) allow a streamer to engage with the audience from multiple perspectives (e.g., the audience could suggest what the streamer should create in the game).

Equipping streamers with hardware is also considered in the literature. Goldberg et al. [90] presented a system in which a human tele-actor, wearing a video camera, headphones, and a microphone, moves through an environment and could take pictures that are uploaded to a remote audience together with a question. An example would be to make a picture of a map and the question "*Which way should I go?*" or a yes-or-no question such as "*Should I open this chamber?*" and here

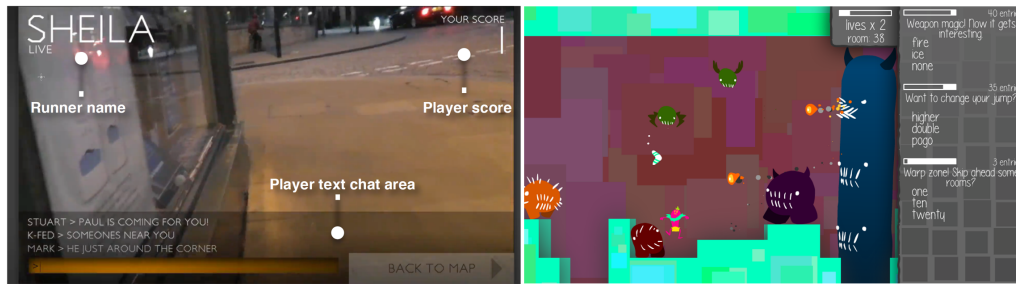


Figure 2.8: Examples of giving the audience more influence. Left: Online view of *I'd Hide You* (taken from [245]). Right: Screenshot of *Choice Chamber* (taken from <https://goo.gl/7zN3i6> (last accessed: 2018-07-07)).

the audience is able to mark points of interests in these picture, i.e., they vote spatially. A similar approach can be seen on *Twitch* today with the so called *Smart Click Maps* extension²¹ showing that such research transfers to the game live-streaming case. Reeves et al. [245] considered *I'd Hide You*, a mixed reality game, in which “streamers” are runners through a city wearing hardware to stream their path while running. They were asked to be entertaining while playing the game. Their goal is to find other runners in the city. An online audience watches the stream and when the streamer comes across an opponent, the audience can take snapshots, which generates points. When the streamer itself is captured by another runner, all audience members of the caught player lose points. The online audience can chat and provide messages to their streamer (see Figure 2.8, left). Thus, overall this mimics the *Twitch* case, but it uses a real-life game. While studies investigated the runner’s perception, they also reason that more coupling to the audience, and further empowering them, would be reasonable (e.g., giving them easier options to provide instructions). *All the Feels* [251] measures the heart rate, the skin conductivity and the streamer’s emotions (through camera data) and visualizes this information to the viewers as a stream overlay. In an “in the wild” study with one streamer, it was found that the system is perceived as useful and can increase the viewer engagement, enjoyment and connection to the streamer. It was also perceived as distracting to a certain extent, leading to the question of how interactions need to be designed to not be distracting.

Overall, these works show that the role of the streamer is important for interactivity. Even without dedicated features, they can show behavior that integrates the audience, for example by simply asking them questions. In addition, the presented approaches also show that giving more information on the streamer itself (such as which emotions are detected) is also perceived positively. Thus a design-space exploration also needs to take into account such aspects. This is something we did in Section 5.2.2, where we also considered viewers’ perceptions of streamers’ behaviors.

²¹ Medium: *The future of live streaming on Twitch is interactive – official launch of extensions*, <https://goo.gl/9T1H7J> (last accessed: 2018-07-07)

Audience-game interactions: Another class of interactions is to allow the audience to have an impact on the game streamed. Today, some games are commercially available that allow viewers to interact. A notable example is *Choice Chamber*²² (see Figure 2.8, right): while the streamer is playing, the audience receives polls. Through them, they can alter the game, for example by deciding which enemies appear or which skills the game's avatar has.

Seering et al. [271] named these kind of experiences *Audience Participation Games* (APGs). They created two APGs (a shooter and a racing game), in which the audience had different options: they could make the game easier by, for example, providing ammo in the shooter game, or teleporting the car ahead; they could make it more difficult by, for example, spawning enemies or mixing up controls; they could decide whether to help or make it more difficult; and they received more insights than streamers, for example where special locations for extra points are. The audience could activate these aspects through entering commands into the *Twitch* chat. Then, either based on a plurality vote (i.e., the command in a time frame that was most often provided was executed) or by simply seeing the command enough times, the change was activated. While the first option requires different viewers agreeing on the command, the second one could be done by one viewer alone (who writes the command enough times). They studied these games in four sessions with one streamer and audiences sizes between five and fifteen. They highlight that APGs provide the viewers with more autonomy. They differentiate between individual agency (i.e., the ability of viewers to affect the game) and social agency (i.e., options to build social bonds with others). Furthermore, they found different viewer characteristics that vary in these two aspects (e.g., viewers that do not engage in having impact on the game). Overall, they emphasize that APGs are a promising design space, that the current lag on the platform is an issue for individual agency, and that “*richer modes of communication, both within games and on the Twitch platform itself, could substantially boost the development of feelings of social agency in game play both through ability to collaborate to achieve a goal and through feelings of commitment to the group*” (p. 436).

Matsuura and Kodama [191] presented a system in which chat messages entered by users are analyzed for their emotional content through a text analysis API. These messages were shown directly in the streamed game (a platform game) and were made a part of it. While the game's avatar can collide with all these messages (providing the player with bonus points or penalties depending on the emotional assessment), the player could also activate special actions by pressing a key while colliding with them. Depending on the positive/negative valence of the message, different in-game effects are carried out (e.g., the player's character receives the option to fly). The authors also modeled an aspect in which a player can only proceed by utilizing this feature. In a small user study the authors found that participants enjoyed this experience, had the feeling that they participated while watching and thought that the system facilitated communication.

²²Created by Studio Bean Games;

<http://www.choicechamber.com> (last accessed: 2018-07-07)

These approaches are interesting, as they show that not only the streaming environment (such as the platform) can be made interactive, but also the content that is streamed. Obviously, while features of the former are usable more generally, i.e., independent of the content streamed, adding features to these games is inherently bound to them. With our considerations of *Helpstone*, we present a system in which viewers have improved communication and interaction channels. Thus, we also considered how to make streaming of a specific game more suitable for the live-streaming context (see Section 5.3), but without giving the audience a direct influence option in the game.

Overall, considering all the presented approaches that investigated aspects of game live-streams, it becomes obvious that empowering streamer-audience interaction is a recent topic. This thesis adds to these ongoing efforts as we consider how to make the live-streaming experiences more interactive and appealing for viewers in Chapter 5. In the next section, we elaborate on shared game control settings that are also of relevance for live-streams where no streamer orchestrates the audience.

2.5.2 Shared Game Control

We first give a general overview on shared control and will then elaborate on its role for the live-streaming context. As a group of users needs to come to a decision, this section also relates back to crowdsourcing (see Section 2.4).

Shared Control in Tasks and Games

Empowering remote audiences for shared control was considered in non-game contexts in the past. For example, in *Apparition* [159] a designer sketches a prototype and describes it via natural language. A crowd, via microtasks and sketch recognition algorithms, translates what the designer has sketched into user interface elements with animation and *Wizard of Oz* [54] style functions. The canvas itself is shared amongst all crowd members. The system provides elements that allow the crowd to self-manage in this shared context, such as an “in-progress” marker to show that a member is currently working on a specific area. Further examples can be seen in the works on crowd input aggregation (see Section 2.4.3), where the crowd for example steered a robot together [160]. As we will see throughout this section, different input aggregations are used in the literature to allow for shared control. Controlling a robot was also investigated in the context of tele-operations. Goldberg et al. [89] allowed participants to control a robot through mouse movements (individual movements were combined). In a test, in which users had to navigate the robot through a maze, it was found that the shared case provided better quality than doing this task alone. In the remainder of this section, we will focus on the shared control of games in general before we move on to the live-streaming case in particular.

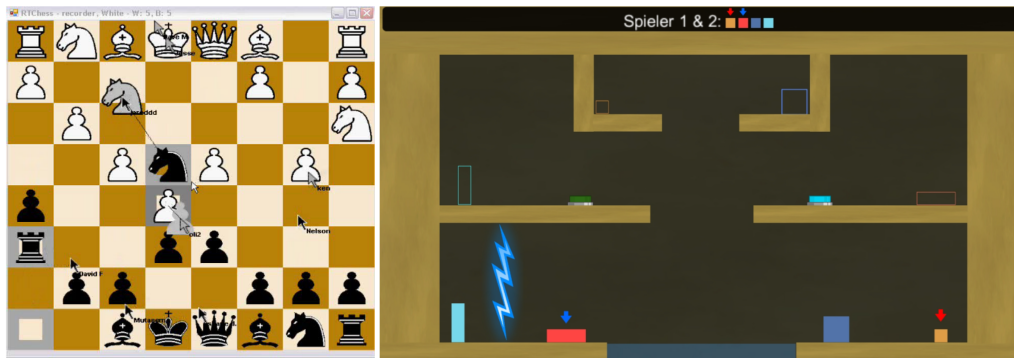


Figure 2.9: Examples of shared game control. Left: Multiple players playing *RTChess* at the same time (taken from [97]). Right: Screenshot of the two-player testbed game used in [69].

With shared game control, we describe the simultaneous control of a game, or parts of it, by sharing a resource. This should be seen in contrast to typical multiplayer games, in which, for example, players play at the same time but have separate resources (e.g., an avatar that they can fully control). Sykownik et al. [291] see shared game control as “an extreme situation in terms of interdependency between players” (p. 848) and “shared control can intuitively be understood as a game control mode, in which players collectively control one single game character” (p. 849). They provided a classification of shared control in which they differentiate the *locus of manipulation (LoM)*, i.e., whether players have separate controls, but share something in the game (distinct *LoM*) or whether they simultaneously control something (mutual *LoM*). As a further dimension they consider the player interdependency (low/high). An example of distinct *LoM* with low player interdependency would be a game in which every player controls an avatar, but can only select one that is not used by another player concurrently. An example of high interdependency, here, would be a game in which all players share the same avatar, but they control different parts of it (arms vs. legs). An example of mutual *LoM* with low player interdependency is a situation in which the control of an avatar is shared in a game by using inputs of a player in a turn-taking fashion (i.e., alternating the controls). A high player interdependency occurs if inputs of all players are processed and are potentially aggregated. For our considerations in this thesis (see Chapter 6), we will focus on the latter aspect.

An example of shared game control is *RTChess*. Gutwin et al. [97] used the basic *chess* game, but changed the rules. Up to 16 players per side are able to play the match simultaneously (see Figure 2.9, left). While the *chess* pieces are only movable based on the normal rules, every piece could be moved at any time and long moves could be intercepted by other pieces. This resulted in a game play which was only limited by the player’s ability to move quickly. No further means for self-administration of the groups were provided here. But still, they found that players start to develop strategies and team coordination occurred, although the game’s speed made this difficult.

Besides this example which could be played online, shared game control has also been considered in a co-located context, as the following works are showing. Battochi et al. [13] investigated a jigsaw-like puzzle game on tabletops with typically-developing children and children with Autism Spectrum Disorders. They added features that enforced collaboration between co-located users, such as that a puzzle piece could only be moved when simultaneously dragged or released by players. In studies they found these specific kinds of shared game control features did not lead to disappointment by any of the child groups. While task completion times (in comparison to conditions in which these features were disabled) were higher, it led to more interactions between the players (e.g., they coordinated and talked more). This exemplifies social aspects attached to shared game control settings, which potentially adds to their appeal, especially given the restricted autonomy (in comparison to typical games).

Fitton and Onyinyechukwu [77] investigated how children in pairs of two (everyone having their own controller) steer a space ship simultaneously. It only reacted when they agreed on the same input (left, right or fire). In two studies the authors found that playing was possible and that strategies (similar to above) were developed to cope with the situation. They found that already in this two-player scenario, disagreement was visible, and also different “play styles” (e.g., sometimes some children were more dominant and told the other player what to do). While the co-location allowed for strategies such as looking at the other controller, and the pairs could easily talk to each other, it is questionable how this transfers to a distributed setting.

Loparev et al. [178] investigated different control schemes for sharing game control for existing video games (such as *Half-Life 2*²³). Their options included approaches in which all players control all aspects of the game (i.e., forwarding all inputs of all players), in which the control of features is split (e.g., one player might do X, while the other can do Y), in which there is turn taking at fixed intervals or random ones (i.e., allowing input of only one of the controllers) and in which different roles are available (e.g., giving hints or playing the game). In addition, the Leader aggregator was used (see Section 2.4.3). In their studies, they found that different strategies were used on top, e.g., that in more difficult game situations novice players give up control to allow more experienced players to handle the situations. Participants reported disliking random changes of turn taking and players, depending on their skill level, had different reactions (e.g., experienced players were often dominant and ignored the hints of others). Overall, though, the authors reported that their participants had fun and were getting more familiar with the concept of shared game control during the study.

Emmerich and Masuch [69] investigated a two-player-platform co-located game (see Figure 2.9, right). They varied whether players had time pressure, whether players were dependent on each other and needed to cooperate or not, and whether players shared the control over the game. Among other results, they

²³ Created by Valve Corporation; see <https://goo.gl/wxbUaC> (last accessed: 2018-07-07)

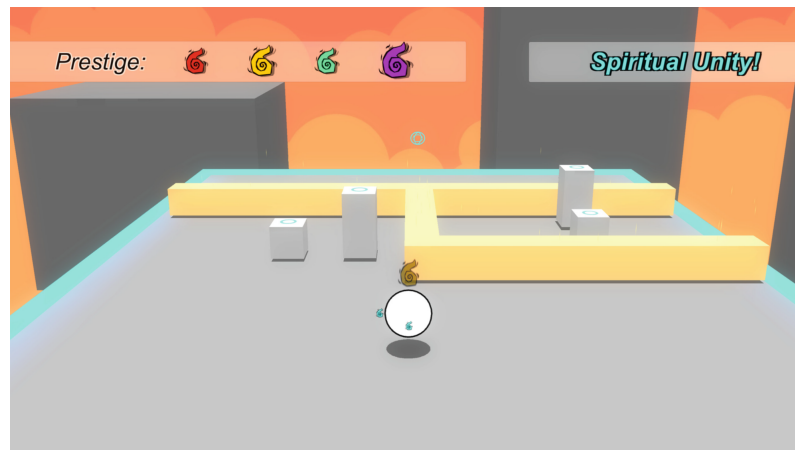


Figure 2.10: The game *Shairit* (taken from [291]).

found that perceived competence and autonomy (even though more options to interact were offered) were significantly lower in the shared game control setting. Compared to normal games where one player has full control, this is not unexpected, and needs to be kept in mind for our considerations as well. Furthermore, participants were less satisfied by their own contributions towards the completion of levels than players in the non-shared control setting. At the same time, the social interactions did not differ compared to the other conditions. It would be interesting to also investigate whether this changes in a live-streaming setting, where people are distributed.

Rozendaal et al. [254] investigated (co-located) shared game control in an *Asteroids*²⁴-like game. In groups of three players, they compared conditions in which each player had full control over a separate spaceship; a condition in which colors of the spaceship could only be changed when the players worked together with their controls (only asteroids with the same colors could be destroyed); and finally, a condition in which the players shared the control over one spaceship and in addition to the aforementioned condition, every player had a different part of the controls (i.e., moving the spaceship, rotating it, shooting). They found that sharing game control affected the levels of experienced sociality, control and engagement. In conditions in which the individual feeling of control and autonomy decreased, the sociality feeling increased, as now cooperation and communication was necessary. The condition in which only the color changing was shared led to the highest level of engagement, as here, both, individual and shared goals were in place. This work underlines that the feeling of sociality can be affected by shared game control in general.

Sykownik et al. [291] developed the game *Shairit* (see Figure 2.10). Here, a sphere needs to be moved through an environment, and needs to collect orbs and move around obstacles. They implemented different options for shared game control: a turn-taking mode in which one player receives full control over the sphere

²⁴Created by Atari; see <https://goo.gl/WzSqh5> (last accessed: 2018-07-07)

in a fixed player sequence for five seconds; and a turn-taking mode in which one player receives full control over the sphere, but with randomized order. Furthermore, the other players can vote to allow the controlling player to be in control longer; a mode in which all player inputs are averaged in time frames and processed into a combined movement direction; and a mode in which all players can control the sphere, requiring collective input. An input processing function increases the movement speed based on the amount of participating players, i.e., if only one of four players interacts, the sphere only moves with one fourth of its speed. In addition, depending on the conformity to the other players, players can activate a mode to receive exclusive control over the sphere for five seconds. They conducted a study in which groups of four co-located players played the game with one of these options each. They found that the game was enjoyable and provided similar results for autonomy, competence and relatedness independent of the shared game mode and that the loss of individual control is not associated with negative experiences per se.

These works show that shared game control is a topic that is currently investigated and that there are mixed results on whether shared game control affects relatedness or the feeling of sociality. In addition, while the shown approaches often tested different “mediation” options, the actual players were not able to select which one they wanted to use for sharing the control. But even then, it became visible that the players develop strategies for how to cope with the situation. Furthermore, it was reported that such experiences impact the perception of autonomy and relatedness, making them interesting overall for this thesis. In the next section, we will focus on shared game control in live-streaming contexts. Here, as a difference from the approaches considered in this section, even more players, which are typically distributed, participate in such shared settings.

Shared Game Control in Live-Streams

In February 2014, the *Twitch Plays Pokémon* (TPP) channel launched on *Twitch* and the game *Pokémon Red*²⁵ was streamed. In this role-playing game, the player’s avatar wanders around, collects creatures and fights against others with those creatures in a turn-based manner. The game’s goal is to win fights against specific non-player characters and to collect all available types of creatures. The novelty of TPP was that no streamer was present and played this game. Instead, the audience played the game simultaneously via entering chat commands (see Figure 2.11) that were retrieved programmatically and mapped to game commands (e.g., typing in *down* would move the avatar or cursor downwards).

Every registered user on *Twitch* could participate in this shared game control setting by joining the channel and entering chat commands. More than 1.1 million people entered 122 million commands²⁶. At the peak, 121,000 people played

²⁵ Created by Game Freak; see <https://goo.gl/ikBKGz> (last accessed: 2018-07-07)

²⁶ Engadget: *Twitch Plays Pokemon final stats: 1.1 million players, 36 million views*, <https://goo.gl/wgoyjn> (last accessed: 2018-07-07)

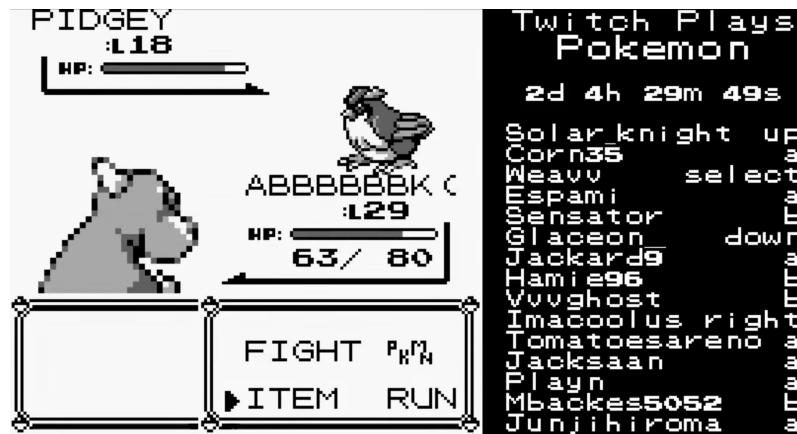


Figure 2.11: Example of *Twitch Plays Pokémon*: A game situation showing a fight and the viewer commands (screenshot taken from the channel <https://www.twitch.tv/twitchplayspokemon> during the first *TPP* run).

simultaneously [154]. The game was finished within 17 days, despite players having different play styles (see below) and despite the live-streaming lag on *Twitch* (see Section 2.5.1). Initially, every user command was carried out. During the play, the anonymous creator introduced a further mode. Here, all entered commands in specific time frames were considered. The mode could be switched by the audience. Thus, this was a means for self-administration, as the audience itself could decide whether everyone should contribute equally (called “*anarchy mode*” in the context of *TPP*) or only the potentially most reasonable command should be carried out (called “*democracy mode*”). *TPP* lived through more than one instance; after the first game was finished it continued successfully with other *Pokémon* games, although not attracting such high viewer numbers [154]. And more *TPP*-like channels appeared with different games, such as playing *Hearthstone* (see Section 5.3) or *Dark Souls*²⁷ (a real-time action role-playing game). But other non-gaming areas were also explored. For example, in *Twitch Installs Arch Linux*, the audience (successfully) installed a *Linux* operating system. *Twitch* also provided its own section for such channels²⁸.

TPP has received scientific attention. It was assessed as a special form of a *Let’s Play* [2], as not only could viewers watch, but they could participate and play themselves, following the “let us play” idea. Such experiences alter the live-streaming experience for the viewers, as they can now participate and should thus have a higher feeling of autonomy. But, based on the nature of the shared game control setting, in comparison to normal games, individual contributions might not become visible anymore, nor might these be carried out in the game, raising the question of how much autonomy a user has in the end [2].

²⁷ Created by From Software; <https://goo.gl/aXoSZF> (last accessed: 2018-07-07)

²⁸ *Twitch* blog: *Announcing the “Twitch Plays” Game Category*, <https://goo.gl/qC52tR> (last accessed: 2018-07-07)

Margel [187] investigated the *TPP* phenomenon by focusing on the occurring social dynamics. He found that these are similar to non-shared-game-control multiplayer settings. In his work he described several noteworthy events that happened during the first playthrough and related these events to trolling behavior [39], the lag issue on the *Twitch* platform and the difficulty of the in-game events in relation to the input aggregation mode changes. Notable events and situations elaborated on by Margel were:

- **The Ledge:** A long and narrow ledge in the game needed to be crossed. When three consecutive *down* commands were provided, the character needed to start from the beginning.
- **The Rock Tunnel:** This area is initially dark until lighted in the normal game. The viewers never lighted the tunnel in *TPP*, and instead they navigated blindly, which took them nine hours.
- **The Spinning Hell:** Viewers needed to navigate through a maze with specific tiles that, when stepped on, move and spin the avatar across the room until a wall is encountered. This led to the introduction of the “*democracy mode*” as progress otherwise was deemed impossible. This section took 26 hours overall. Besides the new input aggregator, a function was also added to repeat specific commands. For example, *up8* would carry out the *up* command eight times.
- **Start9 Protest:** After the introduction of the “*democracy mode*”, players that favored anarchy entered *start9* (leading to an opening and closing of the in-game menu nine times, making further game progress nearly impossible) to express their protest. They saw the new mode undermining the original idea, as the input of many viewers would now have no effect. The voting system for anarchy and democracy was implemented after this protest [2]. Overall, though, the majority of the time anarchy was active [241]. As 80% of the votes needed to be for democracy to activate this aggregator, while only 50% were necessary for anarchy [187, 154], this might be one explanation for this.
- **The Bloody Sunday:** The game allows to manage the captured creature on an in-game computer system. Here, creatures can also be “released” (which effectively deletes them from the game state). In this event, several creatures were deleted as players tried to retrieve one creature, because the “release” option was (accidentally) selected.

These different events were consequences of the aforementioned issues (e.g., *the Ledge* was difficult because of trolls and the *Bloody Sunday* related to the lag issue). Overall, Ramirez et al. [241], highlighted that *TPP* is not comparable to the “*Infinite Monkey Theorem*”²⁹, i.e., it was not the case that the audience finished the game simply because they entered so many commands.

²⁹ Wikipedia: *Infinite monkey theorem*, <https://goo.gl/xtVMyc> (last accessed: 2018-07-07)

Kyriakou [154] analyzed the chat messages during the initial playthrough of *TPP*. Through this, they were able to reveal that many viewers either identified/voted solely for anarchy or solely for democracy, with a smaller number of people voting for both (at least once) during the game. Kyriakou reasons that the modes have similarities to political parties and further underlines that different player objectives were in place. Concerning the content of the chat messages, it became obvious that the “parties” followed different goals, for example making progress vs. entertainment through the randomness. Ramirez et al. [241] further elaborate on the different perceptions of *TPP*. “Anarchists” stated that the game’s appeal was only based on the chaos and thus, it would be less interesting if played “normally”. “Democrats” stated that it is boring if even easy aspects take hours to complete and by using “anarchy”, it would never be able to complete the game [241]. The latter aspect was also underlined by Margel [187]. He identified reasons for when democracy was deemed more valuable by parts of the viewers, such as that overcoming an obstacle would take too long or viewers were aware of the possible damage otherwise (e.g., preventing things such as the *Bloody Sunday*). This shows that there are individual differences in which input aggregation mode is deemed valuable in a shared game control context.

TPP did not happen solely on *Twitch*. Several groups appeared on different social media platforms (e.g., *Reddit*) shortly after the *TPP* experience launched, showing a fandom that appeared alongside it [187]. These platforms were used to discuss strategies, to create tools (e.g., a script that would hide game commands in the chat), to document game progress and to create narratives around the course of action (e.g., based on game events, two fictive religions were created) [2, 154]. While this is a form of self-administration [154], Ramirez et al. [241] even sees this as a meta-game that arose from the actual game, and calls this a “*participatory culture*” (p. 3).

These considerations show that live-streaming is a reasonable context for investigating shared game control. That the different players are not co-located and can join the experience at any time adds further aspects to this setting. Given the *TPP* experience, many social components arose that went beyond the actual game. This fits with live-streaming in general, which was shown to be important from a social perspective as discussed in this chapter. Within this thesis, we also considered shared game control in the live-streaming context: in Chapter 6, we investigated a *TPP*-like setup and provide further aggregators and means for self-administration to investigate whether we can support the players further. Furthermore, in a *chess* context, we analyzed the effectiveness of individual and group decisions. Both aspects add to the current body of knowledge in relation to shared game control.

2.5.3 Summary

In this section, we presented work in the context of live-streaming. This work showed that interactivity and the social experience during streams is an important

driver for the success of live-streams. Although available streaming platforms improved in this respect, there are still shortcomings. While further options are available (e.g., via third-party vendors), it is currently not clear which interaction and communication channels viewers appreciate and how these should be created with the available options. The presented approaches in this section often only considered specific features (for example, those that improve the chat as the primary communication channel) instead of following a general viewpoint. This thesis will add to this in Chapter 5, by investigating which features are used today in channels that are keen on integrating viewers, and how viewers perceive these on a larger scale. Furthermore, we present an approach that provides several communication and interaction channels to integrate viewers more tightly into a streamed game, i.e., we empower them in these contexts. The research on *TPP* has revealed that this is an interesting setting in which to study group dynamics when users receive more options. While the related work, as shown, evaluated different methods for how individual viewer opinions are aggregated, to our knowledge, these have not provided the users similar options as *TPP*, as viewers could not decide which method should be activated to aggregate their inputs. In Chapter 6, we consider further means for self-administration in a live-streaming setting by relying on aggregators known from crowdsourcing (see Section 2.4.3), but will allow viewers to alter which one is active, similar to *TPP*, and derive how these are perceived and used by the players. Considering the Chapters 5 and 6, this thesis thus contributes to current efforts to understand the live-streaming phenomenon and provide insights in how the experience can be further improved, especially in respect to empowering viewers' interactive options.

Chapter 3

Gamified Self-Sustaining Systems

This chapter presents our considerations of systems relying on user participation. Here, the user is not only a consumer of the service, but also an important component to improve its outcome. We developed two prototypes, *ExpenseControl* and the *Trash Game*, which are game-based, crowdsourcing systems, that will be introduced in this chapter. These systems fit our research question **RQ1** (see Section 1.4), as the users have a fundamental influence on these systems themselves. We present user studies that we have conducted in their context, showing that such ideas appeal to users and that gamification is still a reasonable additional motivational layer, even when the tasks to be done have a meaning and benefit for the users.

Section 3.2 is based on the publication [5] and Section 3.3 is based on [166].

3.1 Introduction

As described in Section 2.4, crowdsourcing allows users to exert influence on a system through their contributions when they solve particular tasks. But as discussed in Section 2.4.1, the members of the crowd are often not recruited in the systems where their contributions are used. That means that for the crowd members that solve tasks, it is often not clear how their own contribution relates to a higher cause, or what meaning it has in the end. This, on the other hand, was shown to be beneficial for motivation (see Section 2.3.2 and Section 2.4.1). In consequence, other options to motivate these external crowds are used, especially when the tasks themselves are boring or cumbersome and offer no further personal benefit. In contrast, in this chapter, we consider crowdsourcing systems in which users are affected by their contributions themselves. Here, the users of

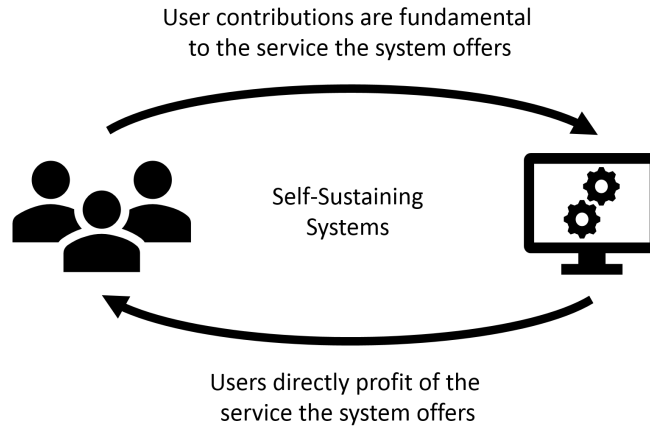


Figure 3.1: Instantiated schematic of reciprocity in self-sustaining systems.

the service are also the members of the crowd that does the crowdsourcing tasks. The system architecture ensures that solving these tasks improves the service outcome. Thus, the perceived influence in such systems should be higher, as users also can directly relate how their contributions impact the system at its runtime. They potentially also experience how the service improves through their participation. Based on this, we see these constructs as self-sustaining systems. While dedicated incentive mechanisms might still be implemented on top in these services, the idea to connect users to the system in this way (i.e., “*My contribution will improve what I will get out of the system*”) should add to the motivation (see Section 2.2.1). These systems are relevant for this thesis as they offer users a fundamental impact on the system’s outcome. Additionally, the systems utilize game-based approaches as a further layer of motivation, which is not uncommon in crowdsourcing (see Section 2.4.1). The latter was done to further motivate the system’s usage, but also the willingness to solve the crowd-based tasks. This also allows us to investigate the role of gamification and game-based aspects in self-sustaining systems, and whether these are able to further impact motivation in such a context. With these systems, we are able to investigate **RQ1** (see Section 1.4). Figure 3.1 shows an instantiated schematic of the reciprocity here. Based on the nature of crowdsourcing, this is a group scenario. Nonetheless, the coupling here is only loose. Through the crowdsourced task nature and through the competition-based aspects, individuals know that they are not alone in improving the system, but could not further interact with each other. Considering the core reciprocity aspect of social systems (i.e., “*My contribution will motivate others to contribute as well*”) [42, 105], known from social networks like *Facebook* or user-generated content platforms like *YouTube* or *Stack Overflow* (see Section 1.2.1), while these might play a role, we do not see these as a necessary component for the self-sustaining systems we consider in this chapter.

We present two game-based self-sustaining systems that are both using image-based microtasks (see Section 2.4.2): *ExpenseControl* (see Section 3.2) and the *Trash Game* (see Section 3.3). Both systems focus on everyday tasks to reach a broad

user base without requiring specific prerequisites to participate. In *ExpenseControl* we focus on the domain of receipt capturing and household accounting books. With it, we were able to validate the idea of a self-sustaining system for tasks that are repetitive and uninteresting in general. We could show that users like the idea of such a system, that their contributions improve the service outcome and that gamification as an additional layer has a strong impact on the amount of work users are willing to perform in such systems. In the *Trash Game*, we considered a self-sustaining system in the context of waste recycling. With it, we were able to show that the system itself also has an impact on the users. The tasks solved not only made the self-sustaining system possible, but also educated users implicitly and playfully; i.e., they can improve in waste recycling by using the system. Overall, both the service and the related studies underline the importance of giving users an option to influence the system's outcome. In the end of the chapter, we will discuss how individuals can be further empowered, especially in relation to approaches we present later in this thesis, such as an option to modify the gamification (see Chapter 4) or to change the aggregation methods at the runtime of the system (see Chapter 6).

3.2 A Self-Sustaining Household Accounting Book

We created *ExpenseControl*, a household accounting book application for mobile devices that allows for tracking expenses by users simply taking pictures of physical receipts. We chose this domain as there is an increasing interest in self-tracking [175, 326] and also a desire to track expenditures, which has been shown to be currently too much effort for many people [140]. Thus, offering a solution for this issue could spark interest in the system itself. This, on the other hand, is a prerequisite for a self-sustaining system, as otherwise, it seems unlikely that users would be interested in putting effort in such a system. After a user has taken the picture of a receipt, the system uses *optical character recognition* (OCR) to extract entities from it, such as the total sum, the store name, individual articles and their prices. In addition, *ExpenseControl* is able to add semantics to entities by categorization aspects (such as that a particular article belongs to groceries). As OCR cannot be considered as perfect (see Section 2.4.2), to enhance the recognition and to derive the semantic information, *ExpenseControl* uses image-based microtasks (see Section 2.4.2 as well) to be solved by its users. The usage of microtasks and the semantic aspects are differences from existing approaches (e.g., rule-based mechanisms [128, 275] or machine learning methods [335]) in this domain. *ExpenseControl* is a self-sustaining system, as these microtasks improve the underlying recognition algorithm and thus the system's service over time. Instead of paying users we use gamification to provide an additional layer of motivation. With *ExpenseControl* we had the following goals:

Goal_{EC} 1 *Creation of a self-sustaining system using non-engaging microtasks:* By developing a system that offers tasks which are not engaging but

improve the system's outcome, the core aspect of the reciprocity in a self-sustaining system can be investigated. The domain of receipt capturing was chosen as context. In *ExpenseControl* basic algorithms were implemented that are able to analyze the receipts and get users *some results*, even if no user would solve microtasks. But satisfying results (based on the nature of OCR; see above) will only be generated through solving image-based microtasks.

Goal_{EC} 2 *Evaluation of the self-sustaining system concept:* By studying *ExpenseControl* we are able to test and validate a self-sustaining system concept in general (i.e., whether microtask solving improves the system's service outcome) and show whether non-engaging tasks are still solved because users know that their own effort will improve the system.

Goal_{EC} 3 *Evaluation of gamification as an additional motivator in a self-sustaining system:* As an additional layer of motivation, *ExpenseControl* offers game elements to motivate app usage and microtask solving. We are interested whether a gamification approach can further motivate participants in a self-sustaining system. Following a non-tailored approach (see Section 2.3), it is also interesting to learn whether individual differences occur in a self-sustaining system setting as well.

3.2.1 Concept and System Design of *ExpenseControl*

The system design and the algorithms of *ExpenseControl* were informed by the results described in an earlier publication by us [140]. Here, we reported on a study in which we analyzed 117 German receipts and an online survey with 238 participants, to learn about requirements and aspects to be considered in a budgeting application. In addition, an informal review of ten budgeting/personal finance apps were conducted to learn about their features. Main aspects that we utilized from this early work are:

- R1** The system should run on a mobile device (e.g., to take the picture directly and easily after receiving the receipt in a store).
- R2** The effort to track expenses needs to be low.
- R3** The application should not only store receipt data, but should also provide statistics on the data stored.
- R4** The results of the receipt analysis informed requirements for the OCR algorithm: that the store's name is hard to extract, that there are multiple synonyms for "total sum" and that it is possible that the article name and price are not on the same line on the receipt. For the latter, [140] also provided seven different common receipt layouts.
- R5** Meta-information is not consistently available on receipts (for example, the category (e.g., groceries) of individual articles).

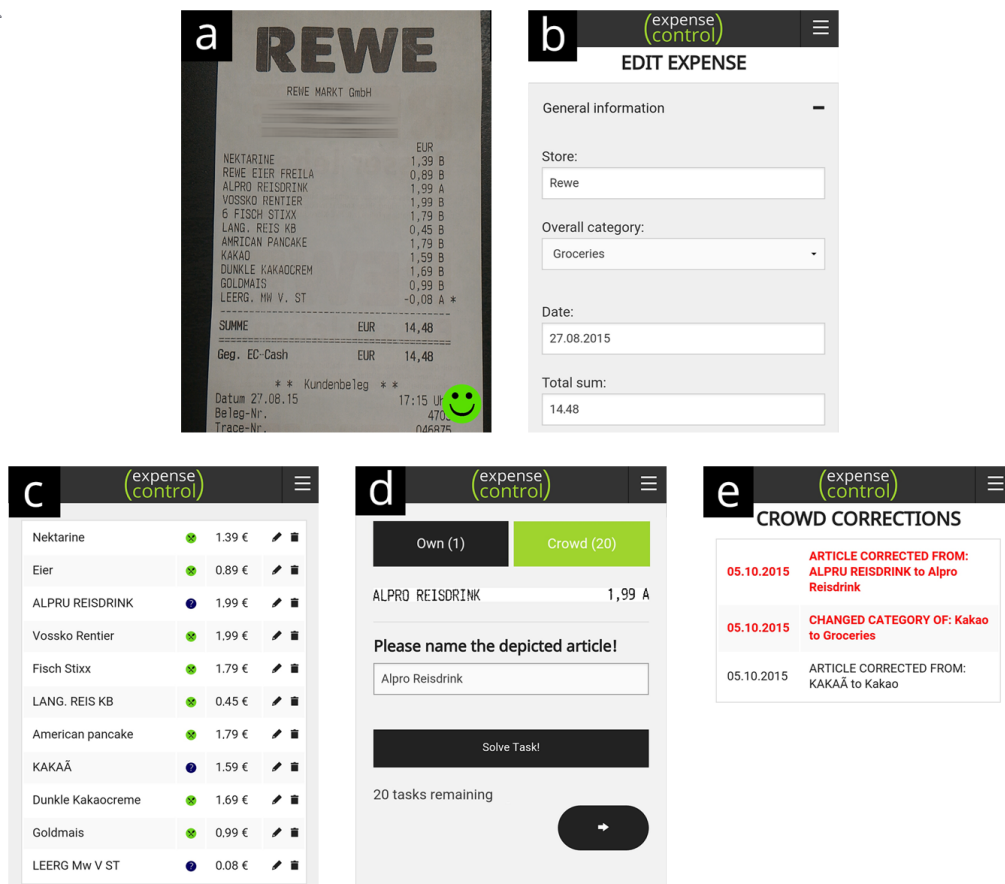


Figure 3.2: Workflow of ExpenseControl: a) Taking a picture. b) Visualization of recognized general information. c) Visualization of extracted articles and their categories. d) Microtask that can be solved. e) Corrections overview.

General Workflow of ExpenseControl

Following **R1**, *ExpenseControl* is designed as a mobile application allowing users to take pictures of receipts (see Figure 3.2a). As soon as a picture is taken, the algorithms (see below for details) start to process the picture and present the result to the user. It consists of the extracted store name, the category of the purchase (e.g., groceries), the date of the purchase, the total sum (see Figure 3.2b) and the extracted individual articles. The latter includes the article name, the category of the article and its price (see Figure 3.2c). A user can now accept what the algorithm has extracted. More effort is not necessary if users want to keep track of their expenses (**R2**). As stated in Section 2.4.2, *OCR* is not perfect, i.e., depending on the receipt paper quality and the quality of the taken picture, errors in these pieces of information are expected, especially when the receipts contain articles that are new to the system. A user has three options to improve the results after he or she has taken a picture: first, a user can manually edit the extracted results. Second, a user can solve microtasks that improve the algorithms and can

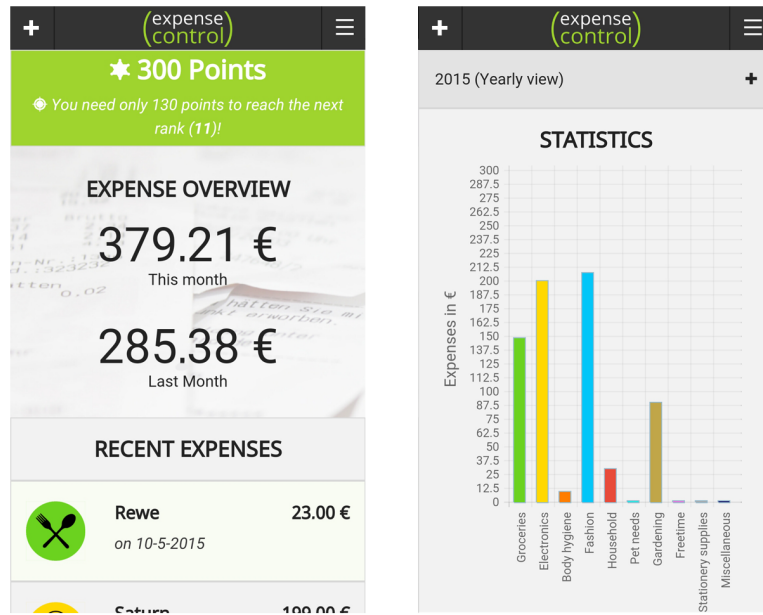


Figure 3.3: ExpenseControl views. Left: Main screen. Right: Statistics view.

decide to solve tasks that relate to issues of his or her own or others' receipts (see Figure 3.2d). Third, a user can wait until others have solved tasks: either their own (which potentially also would contain items of this user) or other tasks (among them, tasks relating to the user's receipt). Both cases lead to updates for this user which are also shown in an overview (see Figure 3.2e).

The receipt data can be inspected by users via the app. Following **R3**, we also provide them with the option to inspect statistics. On the home screen of the application the overall amount of money spent in the current month and in the last month, and the recent expenses, are shown (see Figure 3.3, left). By clicking on the monthly expenses, the user receives an overview on all the details from the corresponding month. *ExpenseControl* also provides a view in which expenses are visualized using different statistics (see Figure 3.3, right) based on categories of single articles (e.g., all expenses for groceries) or the category of the overall expense. A user can also set custom time intervals and can filter expenses. Furthermore, as not all expenses are documented with receipts, a user can add expenses manually.

Algorithms of *ExpenseControl*

Algorithmically, *ExpenseControl* consist of four main parts: receipt capturing, image preprocessing, entity extraction and the microtask improvements. From a system architecture point of view, we decided to preprocess the taken picture by an OCR engine on the user's mobile device to keep the (potential mobile) data traffic as low as possible for the user and reduce the workload on a webserver.

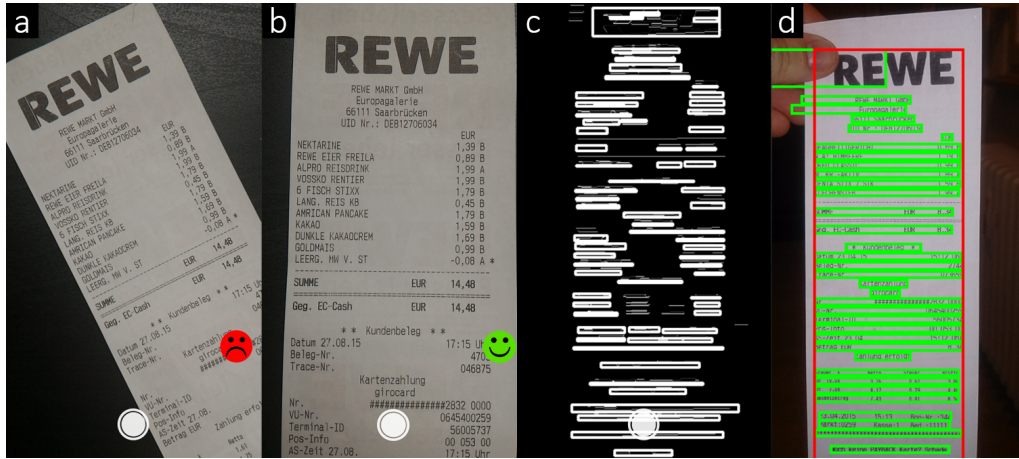


Figure 3.4: Image processing in *ExpenseControl*: a) Unfavorable position of the camera. b) Suitable position to take a picture. c) Identified horizontally aligned regions. d) Regions of interest.

To this server, the OCR result containing recognized text and corresponding line numbers is sent. Here, we extract all relevant information based on the received OCR result using the outcomes of microtasks.

Receipt capturing: As the quality of the picture taken greatly affects the quality of the OCR result [70] and as suggested in [70, 128], we give live feedback in the camera view to raise the chance that the user takes a usable picture (see Figure 3.4a and b). Which kind of emoticon is given depends on lighting and the orientation of the mobile device, since both attributes have been identified as crucial for good OCR results in the literature [70, 114]: with edge detection methods (together with morphological closing transformation), we identify horizontally aligned regions on the images taken. With this, we calculate bounding boxes for these regions (see Figure 3.4c) from which we can conclude whether the device was in an unfavorable position (when boxes are higher than they are wide). Additionally, when the x-coordinates of the bounding boxes are too heterogeneous, we can conclude that there is too much noise in the image, since we found that articles on receipts are always in alignment [140].

Image preprocessing: Images are preprocessed to further enhance the later OCR results [71]: we use the algorithm described above to identify horizontally aligned regions to find regions of interest, i.e., where the receipt is located in the picture. This is done by discarding (for the horizontal consideration only) all horizontally aligned regions where the x-coordinate of the bounding box is either lower than the calculated mean of all x-coordinates, or higher (with a certain threshold). The remaining regions are consolidated to form another bounding box which represents the region where the receipt is located (represented by the red line in Figure 3.4d). This area is then extracted from the picture to obtain better results when thresholding it. Afterward, we used similar approaches (thresholding, Gaussian blur, deskewing) as reported in related work [114, 128].

Entity extraction: After preprocessing we use *Tesseract*³⁰ to extract text regions from the picture, which is then uploaded to a web server. Here, entities, such as the store's name, the total sum and individual articles, are extracted from this. For this, we utilized the theoretical results gained in [140] (R4). Similar to [128], we use regular expressions to identify specific parts of the receipt and divide it into three regions. For the *header* region we look for the first occurrence of a price (everything above belongs to the *header*). For the *body* region, we look for the total sum (several synonyms for this can be found on receipts, so we integrated microtasks to learn these) and the remaining part of the receipt is considered as the *additional information* region. The *header* is further analyzed to extract the name of the store and further information on it, such as the address, the phone number or its URL (again by using regular expressions). The result is compared to our database to identify matching stores, as it might be possible that only parts (e.g., only the phone number) can be extracted. Thus, this look-up would provide related information. The database is extended over time through microtasks to add more store information. The *body* region is then considered to extract article names and corresponding prices. In a first step, all prices are extracted with regular expressions. Next, the algorithm iterates over all lines in the *body*, performs a full text search and uses the *Levenshtein distance* metric [174] for each of them to find a matching entry in our database. In the database, we have corrected versions of entities but also the raw, possibly erroneous OCR version of entities seen and relations to corrected versions. Thus, over time the database will have many OCR versions of the same entity relating to corrected versions and information on it (e.g., where to find the price).

Microtask improvements: The algorithms improve based on the outcome of image-based microtasks (see Section 2.4.2). The idea is that *ExpenseControl*, when used by many users over a longer time span, is able to re-use previous microtask results and thus does not need to generate microtasks for every entity later on. This connection between the algorithms and the microtasks is the core mechanism of the self-sustaining system. Whenever a line from a receipt cannot be classified properly, i.e., the entity cannot be found in our database, a microtask is generated to obtain missing information. In the app, each microtask is shown after another consisting of an image of the unknown receipt line and a short task description. Users can decide to either solve their *own tasks*, which means that these tasks correspond to problems that occurred in their own receipts, or *crowd tasks* that were generated when analyzing receipts of other users. A user can decide to skip a presented microtask because the contained picture may be of low quality or ambiguous in some cases (e.g., if more than one line was extracted accidentally). If a microtask is skipped by at least six users, we discard the picture, similar to [318], as it cannot be used to infer proper information. However, the owner can still manually update the receipt in these cases. We use three different task types to match articles and prices, to extract the total sum (to learn about more synonyms), and to categorize articles (R5) and the overall expense:

³⁰<https://github.com/tesseract-ocr> (last accessed: 2018-07-07)

Figure 3.5: Microtask types in *ExpenseControl*: a) Classification task. b) Article correction task. c) Article categorization task.

Classification task: See Figure 3.5a. This task is generated whenever a line cannot be classified. A user sees the graphical representation of the line and needs to decide what it represents: an article, the total sum or additional information (such as discount/bonus, bottle deposit/return, indication of quantity). Having a classification of a line is fundamental to, for example, match articles to prices.

Article correction task: See Figure 3.5b. This task is generated when an unknown entity is identified as an article. The user is asked to enter the name of the article. The outcome of this task is used to correct OCR spelling errors, distinguish articles and provide a meaningful article name, since the articles are often abbreviated on the receipt. Furthermore, this task also makes it possible to store a relation between the raw text obtained by the OCR engine and the correct version, which helps in future OCR runs.

Article categorization task: See Figure 3.5c. This task is also generated when an unknown entity is identified as an article. In relation to R5, this task provides meta-information, as users need to categorize the article from a list of ten different options (e.g., as groceries).

As soon as a minimum amount of users have participated in a specific microtask, a solution is generated for it. We follow the work of von Ahn et al. [318] and require at least six participants who solved the task as a minimum amount. Depending on the task type, the method to acquire the end result differs. A classification needs to reach at least 60% agreement of all votes. Once this is achieved, the raw OCR result gets stored in our database, together with the determined classification. This allows us to recognize similar entities in the future, provide input for the entity extraction algorithm, and thus decrease the amount of errors over time. The solution for the other task types is based on what the majority of the users stated. The owner of the associated receipt gets notified about all changes and corrections that were performed by the crowd and can be (as stated) inspected in a history (see Figure 3.2e).

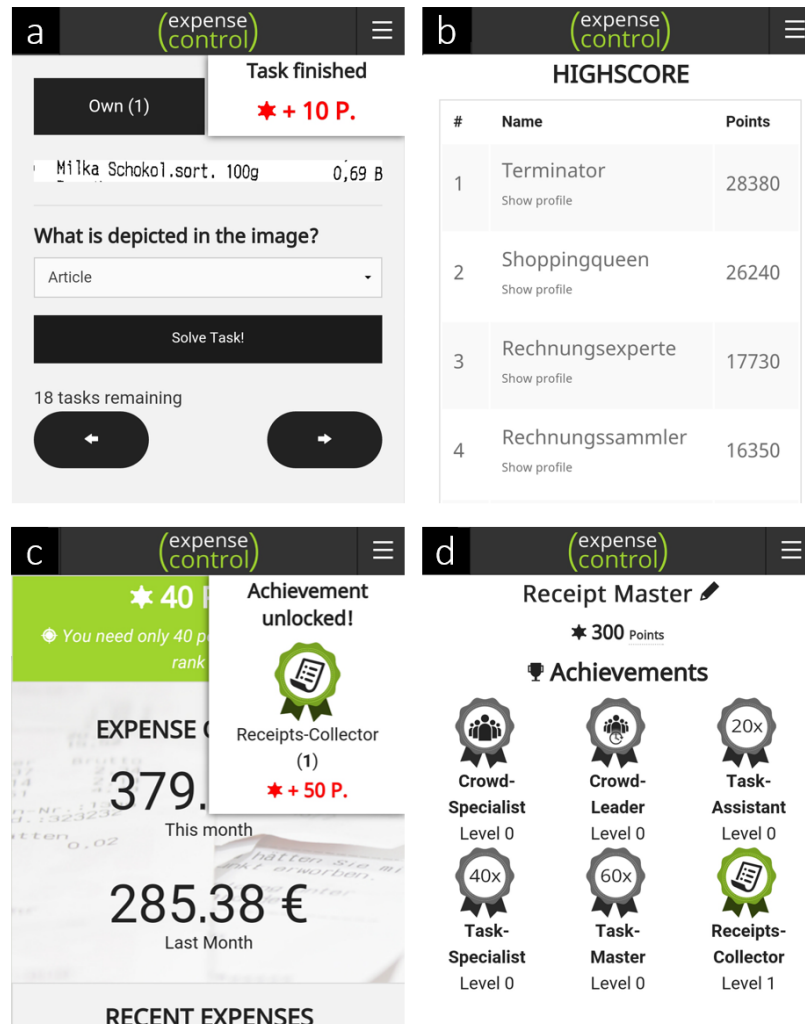


Figure 3.6: Game elements used in *ExpenseControl*: a) Points are awarded. b) The leaderboard. c) An achievement is unlocked. d) The profile view shows all unlocked achievements.

Gamification Approach

We use a gamification approach (see Figure 3.6) in *ExpenseControl* to encourage two aspects. First, we want to motivate users to solve more microtasks. As illustrated, the microtasks will improve the service that is offered and thus are a core part of the self-sustaining system. Second, Snow and Vyas [281] stated that a budgeting app should also use methods to engage users in keeping expenses over a longer time span. Thus, gamification should also motivate people to continuously use the app. Conceptually, this would lead, in consequence, to more receipts, more available microtasks and potentially more active users. Taken together, these aspects should also have positive effects on the self-sustaining system aspect of *ExpenseControl*.

ExpenseControl uses a comparatively simple gamification approach by integrating points, achievements and a leaderboard (which are commonly used in gamification settings [106, 134, 269]). Every task that is solved by a user provides ten points (see Figure 3.6a) and the top scorers are shown on a leaderboard (see Figure 3.6b). The points should not only spark a competitive setting, but we also wanted to give them further meaning, following the idea that users who put effort into the system should be rewarded with more than simply a good position on the leaderboard. If a user solves many tasks belonging to receipts of other users, the system rewards this user by giving microtasks derived from his or her receipts an increased weight. This weight is considered whenever the algorithm selects a new microtask to be presented for any user. High performers thus get rewarded in the sense that their receipts are thereby more quickly corrected, as these are more likely to be presented to others. With this, high performers should receive the feeling that their effort pays off, as the system performs even better for them. In consequence, this might further encourage them to continuously use the app. The amount of collected points and the amount of points necessary to increase the position on the leaderboard are visualized on the main screen (see Figure 3.3a). Alongside the points, we also provide achievements that can be unlocked. These are not only target microtasks, but also the usage of the application. Six different achievements can be unlocked:

- **Crowd-Specialist:** Whenever a user has solved 60 crowd tasks, he or she receives a new level in this achievement and 150 bonus points.
- **Crowd-Leader:** Whenever a user has solved 40 crowd tasks on a day, he or she receives a new level in this achievement and 150 bonus points.
- **Task-Assistant:** Whenever a user has solved 20 tasks on a day, he or she receives a new level in this achievement and 50 bonus points.
- **Task-Specialist:** Whenever a user has solved 40 tasks on a day, he or she receives a new level in this achievement and 110 bonus points.
- **Task-Master:** Whenever a user has solved 60 tasks on a day, he or she receives a new level in this achievement and 160 bonus points.
- **Receipts-Collector:** Whenever a user has scanned 10 receipts, he or she receives a new level in this achievement and 50 bonus points.

Every achievement has a level attached to it, i.e., it can be unlocked multiple times, increasing its level. Achievements also provide points for the overall score as denoted above. Figure 3.6c shows the notification that occurs after unlocking an achievement (which is shown with a corresponding badge). The user has the chance to inspect all unlocked achievements together with their level in the profile view of *ExpenseControl* (see Figure 3.6d). Here, the user can also see the total amount of points. Additionally, the information is provided that solving more of the other's tasks increases the chance that others will solve microtasks belonging to his or her receipts. Users can also inspect how to unlock achievements.

3.2.2 User Study With *ExpenseControl*

We used *ExpenseControl* to investigate whether the idea of a self-sustaining system, with tasks that are not particularly fun, works. We had the following hypotheses:

- H1** *ExpenseControl* subjectively eases keeping track of expenses.
- H2** The outcome of designated microtasks solved by the users can be used to reduce the error rate of captured receipts.
- H3** The outcome of microtasks solved by the users can be used to reduce the error rate of new receipts that are unknown to the system.
- H4** Even in a self-sustaining system context, gamification motivates users to solve more microtasks.

H1 is a prerequisite to assess the self-sustaining system without having a bias and it was the goal we followed with *ExpenseControl* (**R1**, **R2**). **H2** is derived from the system architecture of *ExpenseControl*. As microtasks are generated from captured receipts, results of these can be directly applied to them (as described). As the algorithms use the results of completed microtasks for newly captured receipts, **H3** consequently follows. From a self-sustaining system perspective, **H2** and **H3** are central for **Goal_{EC} 1** and **Goal_{EC} 2**. With **H4** we can validate whether gamification is beneficial in a self-sustaining system, as here, the task solving has an inherent meaning (i.e., the service offered improves) already (**Goal_{EC} 3**).

Method

Participants received the task to use *ExpenseControl* to track their expenses over the course of three weeks. At the beginning, the app was locked until the participant filled out an online questionnaire assessing buying behavior as well as interest in and experiences with tracking expenses (all questions on statements were to be answered on 5-point scales with the labels *disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, *agree*). In the first week, the app was presented without game elements as a baseline to collect how many tasks are solved without any further incentive mechanisms. After this week, the app was locked again until the participants filled out another questionnaire (“pre-gamification questionnaire”) in which we asked questions on the perception of the app parts. In the following two weeks, gamification was activated. We decided against a counterbalanced measures design (i.e., reversing the order of half the participants) since deactivating game elements later could have detrimental effects on participants’ motivation [106]. After the two weeks, participants needed to fill out a post-session questionnaire (“post-gamification questionnaire”), that was similar to the pre-gamification questionnaire, but also included questions on the game elements. Considering our hypotheses, the questionnaires were used to answer **H1** and **H4** from a qualitative point of view. During the study, we stored all receipts that were added and the microtask interactions to investigate **H2**, **H3**

and **H4** quantitatively. To do this, we needed to create a ground truth, i.e., we needed to manually inspect all receipts and classified, categorized and corrected all lines of these receipts. To acquire this ground truth, we followed a two-step procedure: one person provided a ground truth for each line of every receipt that was added by the participants and another person checked for errors (e.g., spelling mistakes or other interpretation options).

Results

In three weeks, 191 receipts were added by twelve participants (seven male, five female; age: 21–30: 9, 31–40: 1, >40: 2). On average, participants added 15.9 receipts (standard deviation $SD=14.4$, median $Mdn=8$). Before the study, only three subjects were keeping track of their expenses, although ten participants claimed to be interested in doing so. Participants reported visiting the same stores (mean $M=4.4$, $SD=.5$, $Mdn=4$) and tended to buy the same products ($M=4.3$, $SD=.5$, $Mdn=4$). Two participants go shopping twice a week, eight of the participants go shopping 3–4 times a week and two claimed to go shopping 5–6 times per week.

Perception of the prototype: Participants stated that the app eases tracking expenses ($M=3.9$, $SD=1$, $Mdn=4$) and that they would rather use our system than manually track their expenses ($M=4$, $SD=1.2$, $Mdn=4$). Moreover, they considered capturing expenses by taking pictures of receipts to be easy ($M=4.5$, $SD=.7$, $Mdn=5$). These findings provide evidence supporting **H1**: the prototype eases keeping track of expenses. The idea that other users could correct data originating from one's own receipts was perceived positively ($M=4.4$, $SD=.7$, $Mdn=4.5$) and considered meaningful ($M=4.5$, $SD=.5$, $Mdn=4.5$). Participants also had the feeling that they used the app often ($M=3.7$, $SD=1.4$, $Mdn=4$).

Entity extraction and crowd performance: 15393 microtasks were solved by the twelve participants, with 1282.8 per participant on average ($SD=1053.5$, $Mdn=1101$). To answer **H2** and **H3**, we calculated an error ratio for each receipt. As a baseline, we considered our algorithm without any crowd input. Since it is impossible to deduce categories for articles in the baseline algorithm (as this is only done through the microtasks), we excluded this aspect for the following considerations. We obtained an error rate by calculating the ratio of the amount of wrong entities (wrong classification and/or wrong value of a line) to the overall amount of entities of a receipt (right classification of a receipt line and correct value) for all receipts. For this, the derived ground truth data was utilized. In the same way, we calculated the error ratio by considering the microtasks (classifications and article corrections) in our algorithm (denoted with crowd-enhanced algorithm (CE) subsequently). The baseline algorithm produced an error rate of 31.8% ($SD=22\%$, $Mdn=32\%$) whereas the CE algorithm reached an error rate of only 10.4% ($SD=14.7\%$, $Mdn=5\%$). A paired t-test showed a significant effect between these error rates ($t(190)=12.5$, $p<.01$) supporting evidence for **H2**: the outcome of designated microtasks solved by a crowd can be used to reduce the error rate of captured receipts.

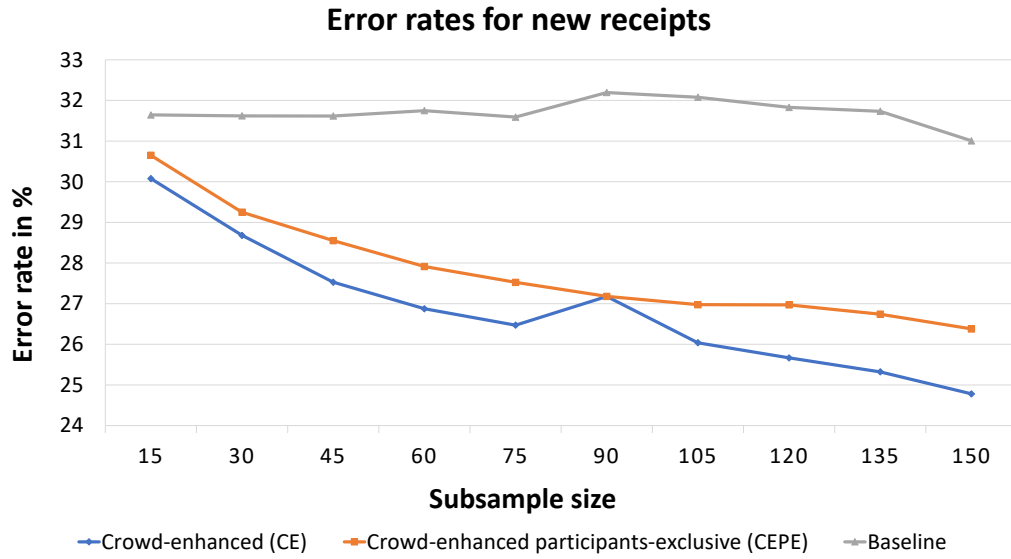


Figure 3.7: The error rate of new receipts for the CE algorithm (blue) and the CEPE algorithm (orange) compared with the baseline algorithm (gray) for different subsample sizes (i.e., how many receipts were used for training).

To evaluate whether microtasks of one receipt are useful for newly added receipts, we picked a random subsample with 15 (30, 45, ...) to 150 receipts and used this as a training set. We then iterated over all receipts that were not contained in this subsample (i.e., the test set) and applied both the CE as well as the baseline algorithm. In the CE algorithm, we only considered solutions of microtasks that were related to unknown entities of receipts within the selected training set to enhance receipts in the test set. Since participants subjectively claimed to buy the same products, we also considered a crowd-enhanced participants-exclusive (CEPE) algorithm in which we used receipts of one user for the test set and the receipts of all other users as the training set to avoid having receipts of the same user in the test and the training sample. To receive more reliable results, we repeated the subsample selection 50 times for each sample size. Figure 3.7 shows the error rates for each subsample for all three algorithms. The results support **H3**: the outcome of microtasks solved by a crowd can be used to reduce the error rate of new receipts that are unknown to the system, since both the CE and the CEPE algorithm outperform the baseline algorithm in all sample sizes. We conducted a repeated measurements ANOVA and found a significant effect between the algorithms ($p < .05$). Pairwise comparisons using the Bonferroni method revealed that the difference between the CE algorithm and the baseline is significant ($p < .01$), as well as the difference between the CEPE algorithm and the baseline ($p < .01$). Moreover, the CE algorithm performed significantly better than the CEPE algorithm ($p < .01$). The fact that the error rate is decreasing with increasing numbers of receipts suggests that our approach improves over time (with increasing data retrieved by the outcome of microtasks). Both **H2** and **H3** show that *ExpenseControl* is a self-sustaining system (**Goal_{EC} 1** and **Goal_{EC} 2**).

Question	Pre-Gamification	Post-Gamification	Significance
Solved many of my own tasks	M=3.6, SD=1.4, Mdn=4	M=3.7, SD=1.6, Mdn=4	p=.72
Solved many crowd tasks	M=4.2, SD=.9, Mdn=4.5	M=4.2, SD=1.3, Mdn=5	p=1
Solving my own tasks was fun	M=2.6, SD=1.3, Mdn=2.5	M=3.2, SD=1.5, Mdn=3	p=.13
Solving crowd tasks was fun	M=2.6, SD=1.3, Mdn=2.5	M=3.4, SD=1.5, Mdn=3.5	t(11)=2.5, p<.05

Table 3.1: Questions and respective answers (on 5-point scales) concerning fun and engagement in the pre- and post-gamification online questionnaire.

Week 1 (No gamification)	Week 2 (Gamification)	Week 3 (Gamification)	Significance
M=238.3, SD=209.9, Mdn=173.5, Min=27, Max=676	M=569.8, SD=462.1, Mdn=522, Min=0, Max=1258	M=474.8, SD=426.7, Mdn=415.5, Min=0, Max=1169	p<.05

Table 3.2: Overview of solved tasks per user/week.

Effects and perception of gamification: To obtain insight about how game elements were perceived subjectively by participants, we asked questions concerning fun and engagement (on a self-created scale) in the pre- and post-gamification questionnaire and compared answers before and after game elements were active in the app (see Table 3.1). In the pre-gamification questionnaire, participants subjectively tended to agree that they solved many of their own microtasks, and agreed to have solved many of those of other users. The answers did not change significantly in the post-gamification questionnaire. However, participants tended to disagree to the statement that solving their own microtasks or solving crowd tasks was fun or engaging in the pre-gamification questionnaire, showing that the tasks are indeed not particularly fun, as we anticipated initially. In the post-gamification questionnaire this perception did not change significantly for their own tasks (even though the mean value increased), but did so for tasks of other users, suggesting that gamification had an effect. All game elements were addressed in the post-gamification questionnaire. The leaderboard was considered most motivating (M=3.8, SD=1.6, Mdn=4.5), followed by points (M=3.7, SD=1.5, Mdn=4) and achievements (M=3.4, SD=1.3, Mdn=3.5).

We considered the amount of solved tasks per user in each week and performed a repeated measurements ANOVA. We found a significant effect between the three weeks (see Table 3.2). Pairwise comparisons using the Bonferroni method showed that the difference between weeks 1 and 2 is significant ($p<.05$) as well as the difference between weeks 1 and 3 ($p<.05$), whereby week 1 was the baseline phase without any gamification elements. These results show evidence for **H4**: the use of gamification additionally motivates users to solve microtasks. Nonetheless, even in the baseline phase participants solved a high number of tasks, adding to

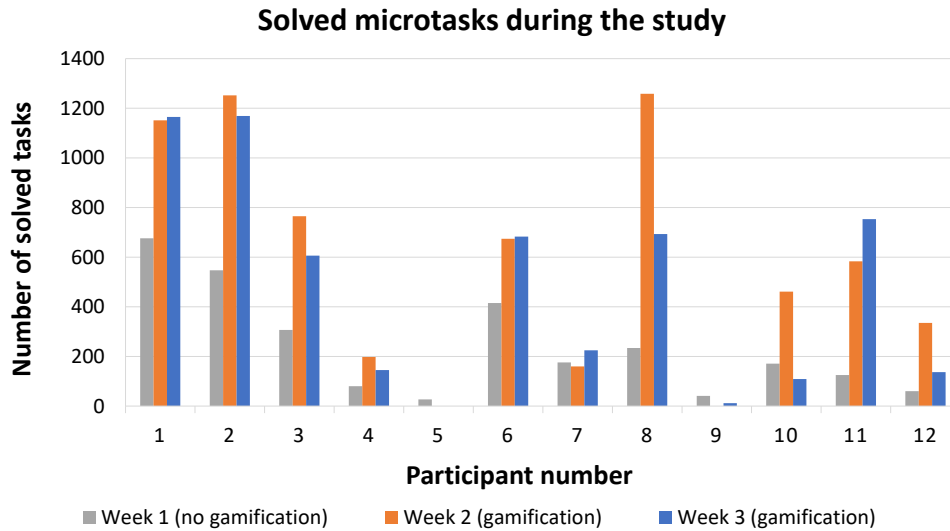


Figure 3.8: Solved microtasks for each participant/week.

the self-sustaining system aspect further. However, the differences between the number of solved microtasks for each participant in every week (see Figure 3.8), and the perception of gamification, indicate that although gamification overall led to a higher number of solved microtasks, there seem to be differences in how gamification is perceived and considered to be motivating. Figure 3.8 indicates that some participants were not affected by gamification at all. It also shows that the number of solved tasks decreases for many participants in the last week again. Although this effect was not significant, it poses the question to what extent the used game elements are able to motivate users in the long run. Works such as [107, 148, 332] have already shown that novelty effects in gamification occur, i.e., over time the attractiveness of the intervention might decrease. Additionally, two participants did not solve many microtasks, showing that neither the gamification nor the meaning of task solving motivated them.

Discussion

The three-week study of *ExpenseControl* revealed positive results as we found supporting evidence for all hypotheses: subjectively, the system eases the tracking of expenses, gamification served as an additional motivator (even in the presence of tasks with a meaning) and the self-sustaining system characteristics could be shown, as not only the recognition performance of entities of existing receipts but also for new receipts improved through the user contributions made in the system. For the latter part, it needs to be kept in mind that even though significant differences were found (between the baseline and our crowd-enhanced algorithm), the factual differences were only a few percent. Based on the results, this may be related to the study design, which had only a small time frame of three weeks and a low number of participants (which can be seen as main

limitation of this study). We reason, that both more time and more users in the system would have a strong impact based on how we designed the algorithms and based on the fact that participants stated that they usually buy the same items. The latter is particularly interesting for *ExpenseControl*, as this means we need many users to receive a broad view on items, but that the system easily improves for an individual. Considering a specific user, he or she will indeed receive the feeling that the system improves (independent of whether he or she or the crowd solved the related microtasks), as the same items more likely have a matching database entry and bad OCR results can be mitigated easily over time.

We also learned that even without gamification, participants are willing to solve a fair amount of tasks, although the microtask solving, as hypothesized, was indeed not perceived as fun. This study cannot state whether the motivation to solve tasks nonetheless comes from the study setting itself based on demand characteristics [195] or the inherent motivation attached to the self-sustaining system and the knowledge that one's own contributions improve it. In favor of the latter explanation is that participants were never explicitly requested to also take part in microtask solving. Instead, their only task was that they should use the app to track expenses, which mimics the *free choice* tasks done in intrinsic motivation research (see Section 2.2.1). What we can see is that gamification increased the amount of solved microtasks considerably and also changed the perception of the tasks (which after the introduction of gamification were perceived as more fun). Consequently, we reason that using gamification in addition is also a reasonable idea in self-sustaining settings.

3.2.3 Contribution to the Thesis' Questions

With the presented system (**Goal_{EC} 1**) and the conducted study we were able to show that the idea of a self-sustaining system is reasonable and works, even when the tasks are repetitive and not engaging (**Goal_{EC} 2**). Even without additional motivation, users participate in it (at least in the study setting) to receive a better outcome of the service offered. The usage of gamification in such settings was shown to be still particularly beneficial, as the amount of solved microtasks more than doubled while gamification was active (**Goal_{EC} 3**). Towards our research questions (see Section 1.4), *ExpenseControl* shows that providing the user with a strong influence and establishing their contributions as a core part of the service is reasonable (adding to **RQ1**; Section 1.4). Moreover, framing such systems with game-based aspects – here by using gamification – is advisable. The study also showed that a “one-size-fits-all” approach is also not optimal in such settings, as not everyone was motivated equally by the chosen motivational approach (i.e., we saw different performances across participants). Chapter 4 presents our investigation of what happens when the user receives options to have influence on the gamification itself. In the next section, we consider a self-sustaining system which, in contrast to *ExpenseControl*, also provided feedback on task solving, allowing us to investigate its impact on users in this context.



Figure 3.9: Examples of augmented trash cans. Left: A can showing its interior (taken from [230]). Right: The *BinCam* (extracted from [46]).

3.3 A Self-Sustaining System to Improve Recycling

The topic of encouraging people to reflect on their recycling behavior (e.g., [336]) and waste in general (e.g. [94]) has been under investigation for years now. This seems reasonable given that world cities in 2012 generated about 1.3 billion tonnes of solid waste per year, which is expected to increase to 2.2 billion tonnes by 2025 [116]. In terms of recycling, in Germany, different trash bins for households are available that are designated to hold only a specific kind of trash, and in addition, glass containers (different bins for clear, green and brown glass) for non-returnable bottles can be found in all neighborhoods [72]. If the separation of garbage is done properly, it has a positive effect in terms of environmental protection, e.g., by reducing CO₂ emissions [72]. Zlatow and Kelliher [336] highlight that almost 70% of the material contained in recycling bins is deemed unusable due to contamination from elements that do not belong in it. Two reasons, amongst several, are the complexity of the recycling rules and lack of motivation [47, 117, 313, 325]. This topic is also considered within HCI. For example, Reif et al. [247] highlight the importance of direct feedback and propose to augment public trash cans with technology (e.g., displaying relevant information on the trash can screens) and Paulos and Jenkins [230] discussed a public trash can which is able to recognize when trash is added/removed. An integrated projector shows its interior contents in front of it (see Figure 3.9, left). The authors envisioned a direct interaction with the trash can by allowing individuals to send text messages, which would then also be displayed on the street. Both approaches had the goal to encourage people to reflect on their behavior and/or to educate people on how to recycle correctly.

Another example of an augmented trash can that also integrated a crowd-based image classification approach is the *BinCam* system [46, 48, 299], which has received broad media coverage [47]. Kitchen bins of shared houses (to reduce privacy issues) were augmented (see Figure 3.9, right) and were able to recognize when new objects were discarded. Then a picture was uploaded to a *Facebook*

page on which other *BinCam* households were also active. The goal was to raise awareness and to subsequently support intentions for behavior changes [46]. Besides the social pressure due to other people on *Facebook* being able to see which items were potentially incorrectly discarded, the authors also added some gamification elements (i.e., a form of points and achievements could be gained, to be competitive in relation to other households). The latter aimed at engaging people further in not adding objects that would not belong in a given trash can. The pictures taken were also sent to *Amazon Mechanical Turk*, where the paid crowd had to count the visible objects in terms of the actual waste category they belonged to (being an image-based microtask). This information was then used in the gamification approach and shown as picture tags on *Facebook*. The performance of the crowd was problematic, as in a random picture sample of 20 pictures, only five were correct [299]. For our second self-sustaining system, the *Trash Game*, we assessed this context as interesting and used the *BinCam* idea as a basis, but changed the following major aspects: first, instead of working with shared houses, we decided to use a public trash can with several bins to build a recycling game out of this and to minimize privacy concerns further [46, 299]. Second, instead of using an external, paid crowd that carries out the image classification task to identify the waste category of the objects, we let the users of our system carry out this task with a gamified component connected to the trash can, to investigate the self-sustaining system aspects. Third, in contrast to the *BinCam*, we provide feedback on the classification. These aspects are an opportunity to educate on a larger scale and allow us to validate whether feedback in a self-sustaining concept has a further impact on users. Consequently, we had the following goals, which are targeted and discussed in the remainder of this chapter:

- Goal_{TG} 1** *Investigation of a feedback loop in image-based microtasks:* Individuals might, while classifying pictures of waste, also be affected by the task and adapt their behavior subsequently when receiving feedback to their task solving. Thus, we can learn whether we can educate implicitly on a larger scale when targeting the users of self-sustaining systems.
- Goal_{TG} 2** *Creation of a gamified self-sustaining system with a feedback loop:* In contrast to *ExpenseControl* above, in which no feedback was given on the individual task solution (i.e., for a user it remained unclear whether his or her solution was correct or not), we assumed that giving feedback makes the tasks more interesting, and based on the previous goal, will affect users. To target this goal, we create a system that makes use of a feedback loop to show how this can be used in the self-sustaining system context.

3.3.1 Studying the Wisdom of Crowds in Waste Recycling

Towards **Goal_{TC} 1**, we used an online questionnaire with which we studied individual capabilities in classifying waste, participants attitudes' towards waste separation and whether they improve in this task over time while classifying. Additionally, we wanted to validate whether the *wisdom of crowds* [289] effect is indeed not applicable in this setting (as the paid crowd in the *BinCam* approach did not perform well [299]). These aspects are important to know before a self-sustaining system in the context of waste recycling is built. Thus, this study served as a foundation. More specifically, with the study we tried to find evidence for the following hypotheses:

- H1** An individual is in general not capable of classifying waste without errors.
- H2** Aggregating the individual classification results leads to lower error rates in comparison to only considering individual decisions.
- H3** Individuals improve when receiving feedback on recycling decisions.
- H4** Showing the results of other users positively impacts individual performance further.

H1 is motivated by related work showing that people have problems classifying waste correctly (see above). **H2** is based on the idea of the *wisdom of crowds* [289]. **H3** focuses on the educational aspect: in [302] it was shown that feedback (on incorrectly separated objects) based on an analysis after the garbage had been picked up had a positive impact on future decisions. We want to replicate this effect in our setting, as this is important for the implicit educating aspect of the self-sustaining system later on. **H4** is based on the idea that a comparison to peer decisions might have a stronger impact than right-or-wrong feedback alone. While in a general Q&A game context [193] no effects for different kinds of feedback were found, this might be different in this context.

Method

We implemented a gamified online questionnaire consisting of behavioral questions, classification tasks and a retest of false classifications. The classification process mimics the process we envisioned for the users in a potential self-sustaining system later on. The gamification elements were meant to encourage people to finish the questionnaire, but were also used to assess potential elements for the system realization later as well: we added points if a classification was correct, subtracted points if not and provided bonus points if an answer was given quickly and correctly. During the classification tasks (depending on the group; see below) the participants could see how many points they needed to get to the next place on the leaderboard and how big the distance from the previous position on the board was. In the end, the participants were shown the complete leaderboard, on which only nicknames and points were displayed.

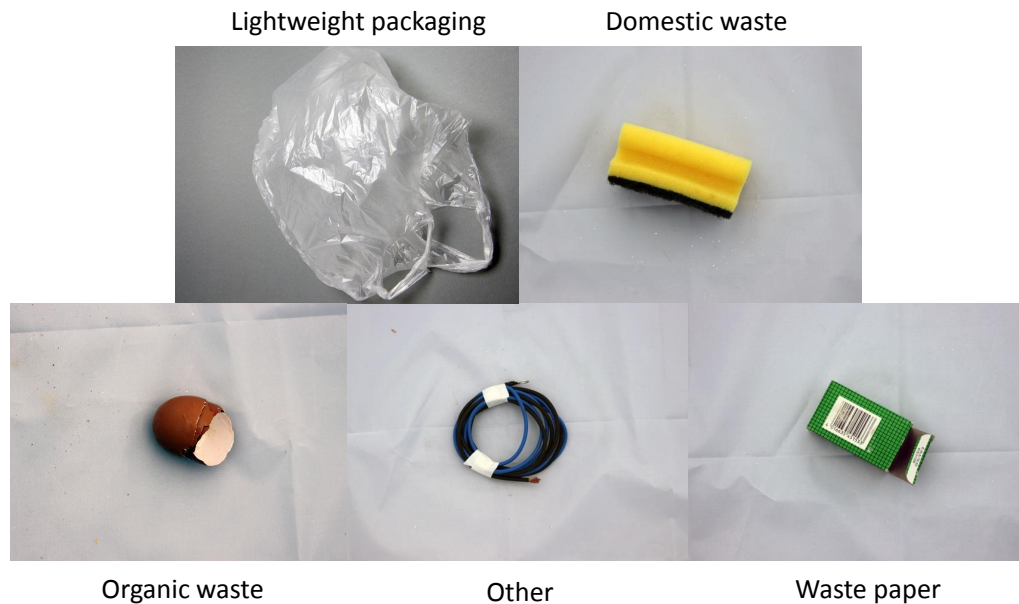



Figure 3.10: Example pictures for every considered waste category.

Behavioral questions: After the introduction, in which we explained that the goal was to receive insights into waste separation behavior of German people, we asked the participants whether they think waste separation is easy, whether they are able to separate waste correctly and how complicated they judge waste separation to be in Germany. These questions (and all the other behavioral questions) were to be answered on a 7-point scale, with the labels *strongly disagree* and *strongly agree* shown on the extreme values. A set of classification tasks (see below) followed and subsequently we let participants judge their own performance in this task overall and continued with questions on the game elements used. After that, we asked questions about their waste separation behavior and collected demographic data.

Classification task: We informed participants that pictures of objects (see Figure 3.10 for examples) would be shown and that they needed to state how they would be disposed of correctly in Germany (domestic waste, lightweight packaging, waste paper, organic waste or none of those categories), we would assign points for their answer and the faster they answered, the better for their overall score. To acquire a ground truth, we consulted official material on the topic of how to separate garbage in Germany correctly and selected only objects for which we found consensus in it. Every picture was rated by three judges beforehand in terms of its difficulty level and in case of incongruity, a discussion solved it (in five cases such a discussion was necessary). The selection of objects was guided by the goal to have different representatives per difficulty category (three easy pictures, three medium pictures and two difficult pictures per category), resulting in $8 \times 5 = 40$ pictures. An example of an easy task was an apple (organic waste), one for a medium task was a non-returnable can (lightweight packaging), and a

THE TROSH GAME



Continue >>

Skip >>

How would you dispose of the shown object?

Waste paper							
Light-weight pack.				X			
Domestic waste							
Organic waste							
Other							
	1	2	3	4	5	6	7
	Not confident Very confident						

Comment:

Figure 3.11: Example classification task of the questionnaire.

supermarket receipt (which consists of thermal paper that should be disposed of as domestic waste) was assessed as hard to classify. Unlike the *BinCam* [299], we planned to use computer graphics algorithms later on to extract one disposed object at a time; hence, every picture contained only one object.

Participants could skip classifications completely and could add a comment to each of them. In addition, after they received feedback (depending on their group; see below), they also had the chance to comment on it. For every decision, participants also had to indicate how confident they were in it on a 7-point scale. To simplify this, we used a grid layout in which the participants could assign a classification answer and a confidence with only one click (see Figure 3.11). We also measured the time per answer as an unusually long time might be an indicator that the participant used outside information. The order of the pictures was randomized and participants were assigned to one of five groups randomly (but equally) at the start. We varied the feedback participants received after a classification, depending on the group:

- **No feedback (*NF*):** Participants do not receive feedback and only see their score and the leaderboard after all tasks. This condition serves as a baseline.
- **Ground truth feedback (*GT_FOnly*):** Participants always see whether their decision was correct or incorrect, and the correct answer. All gamification elements are available, i.e., they can see their points, their current placement on the leaderboard and how many points they are from the next position.

- **Ground truth feedback with explanation ($GT F_{Explanation}$):** Same as $GT F_{Only}$. In addition, a short statement and a link to an official document explained the ground truth.
- **Ground truth feedback with same crowd decision ($GT F_{CrowdSingle}$):** Same as $GT F_{Only}$. In addition, they see how many people decided in the same way, by seeing a percentage (e.g., 12% had the same opinion).
- **Ground truth feedback with crowd decisions ($GT F_{CrowdAll}$):** Same as $GT F_{Only}$. In addition, they see how the crowd decided, by seeing a percentage per classification option.

Retest of false classifications: To check if any improvements were achieved through the feedback, we asked the participants whether they wanted to improve their score by classifying more pictures before seeing the leaderboard (“bonus run”). These pictures were selected based on the errors made. The feedback group assignment remained the same for this task. In addition, we asked whether they wanted to receive an additional invitation to a follow-up study. One week after completion of the questionnaire, they received an e-mail with a new link. The new questionnaire consisted of two questions (“*Did you consider the topic of waste separation more during the last week?*” and “*Do you think that you disposed of and sorted waste more thoughtfully during the last week?*”) and a set of classification tasks split in two chunks. The selection of the objects was again guided by the errors a participant made in the original questionnaire. In the first chunk, instead of pictures we showed only the name of the object (everything else in the task remained the same). This should have minimized picture recognition effects. In the second chunk, we used the same pictures again. The order of text/pictures in both chunks was randomized and no feedback was provided.

We set up the questionnaire in German and required participants to have lived at least three years in Germany to participate. This was done to increase the chance of participants being familiar with the German recycling rules. The questionnaire was published on student mailing lists and social networks.

Results

184 people completed the questionnaire (93 male, 78 female, 13 no answer; age: <21: 29, 21–30: 124, 31–40: 12, 41–50: 12, >50: 7). Mainly, 87 participants (47%) reported being students and 59 (32%) being employed.

Waste separation behavior: People tended to agree to that waste separation is easy ($M=5.1$, $SD=1.3$, $Mdn=5$), that they are able to separate waste correctly ($M=5.2$, $SD=1.3$, $Mdn=5$), that they are eco-sensitive ($M=5.1$, $SD=1.4$, $Mdn=5$), that they separate to the best of their knowledge ($M=5.4$, $SD=1.6$, $Mdn=6$) and that waste separation is important to them ($M=5$, $SD=1.4$, $Mdn=5$). The 54 participants (29%) who disagreed (selecting four or less on the scale) were shown potential reasons: 17% selected that it is too complicated, 30% that it is too much effort,

	Paper	Lightweight	Domestic	Organic	Other
Easy	5.5%	4.2%	24.1%	1.7%	8.1%
Medium	11.6%	11%	45.2%	21.1%	24.2%
Difficult	25.4%	31.1%	87.8%	20.1%	68%
Overall error	12.8%	13.5%	48%	13.5%	29.1%

Table 3.3: Errors per waste category and a priori assigned difficulty level, aggregated over all participants.

	Paper	Lightweight	Domestic	Organic	Other	Skipped
Paper	87.3%	4.8%	6.2%	0.6%	0.9%	0.3%
Lightweight	0.8%	86.6%	9.6%	0%	2.8%	0.1%
Domestic	22.4%	13%	51.9%	1.7%	10.6%	0.6%
Organic	1.3%	0.1%	11.2%	86.5%	0.6%	0.2%
Other	0.1%	11.6%	17.2%	0%	70.8%	0.3%

Table 3.4: Confusion matrix showing the aggregated classification results (paper waste, lightweight packaging, domestic waste, organic waste, other).

39% that there is too little space for waste separation at home, 31% that waste separation is useless and 26% stated that incentives are lacking. In contrast, of the 130 participants who at least somewhat agreed (71%), 89% want to save resources and protect the environment, 35% have financial reasons and 65% also stated that it is their responsibility to protect the environment. The participants tended to disagree with the statement that they seek information if they are unsure how to dispose of waste correctly ($M=3.2$, $SD=1.6$, $Mdn=3$) and that waste separation in Germany is difficult ($M=3.5$, $SD=1.6$, $Mdn=3$). After the classification task, we asked again whether they are able to separate waste and observed a small, but significant decrease in the self-assessment with respect to their capabilities in waste separation ($M=4.9$, $SD=1.2$, $Mdn=5$, t -test: $t(183)=3.4$, $p<.01$, $d=.25$).

Performance in the classification tasks: We deleted answers in the classification tasks that took longer than three times the average answering time [197] ($M=7s$, $SD=104s$) as an explanation for this might be that participants had utilized external material. In the end, 7236 classifications in the main run were considered. Skipped classifications (22) were counted as wrong classifications. On average an individual made $M=23.4\%$ errors ($SD=7.6\%$, $Mdn=22.5\%$), which supports **H1** – an individual is not able to classify waste without errors in general. We analyzed the errors per waste category and with respect to our a priori assigned difficulty level. Table 3.3 shows that the a priori levels fit in general (with organic waste as the only exception), i.e., medium and hard labels produced more errors than objects rated easy, and showed that some objects are harder to dispose of correctly. We analyzed the kinds of errors made more deeply, i.e., if an error was made, we checked which other category was selected. The corresponding confusion matrix is shown in Table 3.4. Considering the difficulty levels, it shows that the

domestic waste objects produced the most errors. Another aspect which is also of interest is the relationship between assigned confidence value and the actual decision. We found that many participants utilized only the extreme values (a reason might be the time aspect to achieve more points). For the subsequent analysis, participants were excluded based on the standard deviation ($SD_{Confidence} < 0.1$) or the mean values (if not in $1.5 \leq M_{Confidence} \leq 6.5$) of their confidence scores overall: the remaining 88 answers were used for analysis. It showed that if people decided wrongly, they provided on average a lower confidence score than if they decided correctly ($M_{incorrect}=4.9$, $SD_{incorrect}=1$, $M_{correct}=5.7$, $SD_{correct}=1.1$, $t(87)=11.2$, $p<.01$, $d=.79$).

Assessing improvements over time: Participants were able to improve:

During the run: After removing outliers, we compared the error rate in the *GTF*-conditions ($M=22.6\%$, $SD=6\%$) with the *NF*-condition ($M=25.2\%$, $SD=8.8\%$), but a t-test showed no significant difference ($p=.5$).

Bonus run: 123 participants (66%) completed the bonus run. We expected an improvement in all *GTF*-conditions, based on the recognition of the pictures and remembering the correct answers. As the option to receive bonus points was not articulated before the classification task and answering the behavioral questions served as a distractor, a better performance indicates that participants had seen the correct result and could also remember it later on. The average error rate of the users here was $M=80\%$, $SD=21\%$ in the *NF*-condition and $M=12\%$, $SD=13.1\%$ in the *GTF*-conditions. A one-way ANOVA showed a significant effect between condition and error rate (Welch's $F(4,77.5)=68.1$, $p<.01$, est. $\omega=.83$). The Games-Howell post hoc procedure was used since the homogeneity of variance assumption was not met, and revealed that every *GTF*-condition is significantly different from the *NF*-condition (every comparison with $p<.01$). We also checked whether the crowd feedback had any impact on the way users performed. However, none of the comparisons showed a significant effect.

Follow-up run: As the bonus point run could still have only a short-term effect, we wanted to measure whether we could find evidence for improvements after a week. As a limitation, only 36 participants (*NF*: 12, *GTF*: 5, *GTF_{Explanation}*: 3, *GTF_{CrowdSingle}*: 8, *GTF_{CrowdAll}*: 8) took part in this second study, limiting its expressiveness. We asked whether participants sought information about waste separation between the two questionnaires ($M=2.9$, $SD=1.8$, $Mdn=2$) and whether they separated waste more conscientiously ($M=2.9$, $SD=1.7$, $Mdn=3$), but they tended to disagree to both. As outlined, we had two chunks of the classification task. The average error rate of users previously in the *NF*-condition in chunk 1 (chunk 2) was $M=54.7\%$, $SD=11.4\%$ ($M=54.3\%$, $SD=14.3\%$) and of users previously in the *GTF*-conditions was $M=33\%$, $SD=18.2\%$ ($M=23.8\%$, $SD=6.6\%$). Because of the low number of participants in *GTF_{Explanation}*, we excluded this condition in the one-way ANOVA analysis of chunk 1 and 2; additionally, one participant completed only chunk 1. We found a significant effect between group and error rate in each chunk ($F(3,29)=5.6$, $p<.01$, $\omega=.54$ and $F(3,28)=26.3$, $p<.01$, $\omega=.84$).

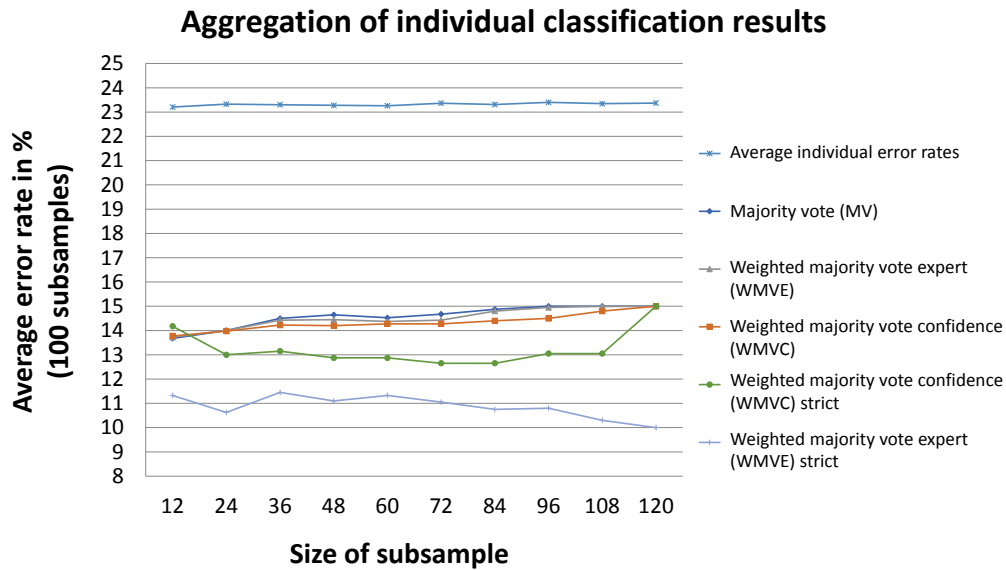


Figure 3.12: Crowd error rate in relation to aggregation algorithm and averaged individual error rates for different subsamples.

Gabriel’s post hoc procedure revealed that every *GTF*-condition is significantly different from the *NF*-condition (for chunk 1 with $p < .05$, for chunk 2 with $p < .01$), indicating that feedback was indeed helpful for subsequent classifications, even after a week. The error rates over all conditions in chunk 2 ($M = 34.2\%$, $SD = 17.6\%$) compared to chunk 1 ($M = 38.5\%$, $SD = 16.4\%$) were significantly lower ($t(34) = 3.7$, $p < .01$, $d = .27$) indicating that people indeed utilized recall effects. In both chunks feedback had improved the user performance, providing evidence supporting **H3**. Again, no effect could be found by considering the crowd conditions. Hence, **H4** cannot be supported with the results found.

Crowd performance: We tested different aggregation methods (see Section 2.4.3) to see if the crowd produces better results than individuals in this setting: a simple majority vote (MV); a weighted majority vote taking the provided confidence scores into account as weights (WMVC) and a weighted majority vote considering the percentage of correct decisions as weights to identify experts (WMVE). The problems with the confidence values (see above) resulted in a strict condition, in which we set the weights to zero for those who apparently did not use the confidence value properly. We also integrated a strict value for the expert rating (setting the weight to zero for participants having a higher than average error rate). To receive more reliable results we took random subsamples of the dataset and repeated the subsample selection 100 times. Figure 3.12 shows the aggregated results in terms of error rates. For a better direct comparison, we also averaged over the individual (aggregated) error rate in the corresponding subsample. The result supports **H2**: the *wisdom of crowds* is applicable, as (even a small) group produces a better result, independent of the aggregator (although the strict expert metric worked best).

Perception of feedback and gamification elements: For the assessment of the perception of feedback (namely true/false feedback, seeing the correct answer, justification for the ground truth answer and how the crowd has decided) and gamification elements (seeing points, position on the leaderboard, competition in general) we asked whether the element motivated them to classify more pictures (A), or if the participants belong to a condition in which the element was not used (B), whether it would have motivated them. The feedback elements used were seen as motivating (Mdn=5 for the different aspects in A, Mdn=6 in B); the only exception was the crowd feedback (A: $n=74$, $M=3.5$, $SD=2$, Mdn=3, B: $n=110$, $M=3.7$, $SD=1.8$, Mdn=4), with no significant difference between $GTFCrowdSingle$ and $GTFCrowdAll$. An explanation could be that such feedback is not interesting when the correct answer is also shown. Here, a follow-up study needs to investigate this further. The gamification elements received mixed results (Mdn=4 for seeing points, other Mdn=5). In all but one case participants not exposed to the corresponding element provided slightly higher scores. Always seeing one's current position on the leaderboard is the only aspect which was perceived better by participants exposed to it (A: $n=148$, $M=4.2$, $SD=2.2$, Mdn=5, B: $n=36$, $M=3.5$, $SD=1.8$, Mdn=4). Overall, the mixed answers also underline that there are individual differences in the perception of game elements, fitting to the considerations described in Section 2.3.

Discussion

The study showed that German-speaking individuals make errors in recycling although they reported being eco-aware and capable of separating waste correctly. The latter could also be explainable by a social desirability bias [215] in this study setting. Nonetheless, overall this supports **H1**, serving as a basis to motivate the search for proper user assistance in this task. The conducted analysis showed that the *wisdom of crowds* is applicable in this context, as considering the group of users performed better than the individual (**H2**). Depending on the aggregation algorithm, the group produces roughly only half as many errors as an individual on average. In contrast to the *BinCam* [299], the group in our study performed better. Reasons for that might be that congruency of nationality and waste disposal rules were ensured and that only one object should be classified at a time instead of counting all elements in a bin (which might be hard, as objects might be partly covered by objects above them). In our case, participants were not paid, which might have had an influence in that they did not just select something to gain the money fast. Instead maybe they decided more thoroughly based on the feedback and the game. We found evidence supporting **H3**: participants used the feedback shown and produced fewer errors, not only for the same pictures, but also for text-only representations after one week and without knowing that they would be re-tested. Considering that the game elements introduced a time factor, it might even be the case that participants misjudged when time was counted and skipped through the feedback, instead of reading it. If this is true,

the effects in the feedback conditions might currently even be underestimated. Finally, the gamification and feedback elements were perceived positively. With this setup, though, we could not find supporting evidence for **H4**, indicating that the feedback framing has no influence.

These results show that a self-sustaining system is reasonable to use in this setting: a crowd-based approach seems in general possible and by providing feedback to the users, they are likely to improve implicitly over time, showing that such systems also have effects that go beyond actual task solving. The results indicate that feedback in such an image-based microtask helps people even when they are unaware of being retested (**Goal_{TG} 1**).

3.3.2 Concept and System Design of the *Trash Game*

In this section, we present the *Trash Game* a self-sustaining system consisting of two components: a modified public trash can with cameras in bins for different types of waste (i.e., users need to decide where to insert their garbage) and a mobile app. While the trash can produces images from discarded objects (and is able to derive which bin was used), the mobile app presents them and integrates classification tasks in its design. Both systems present feedback on whether the classification decision was correct or not. Thus, both aim at educating the users in the overall system: the local group of users using the trash can to dispose of something correctly and the distributed group of users playing the mobile game. Furthermore, both components have the goal to get people engaged in sorting correctly. The trash can by displaying whether the object was inserted into the correct bin and building a competition around it, and the mobile app by motivating people to improve their fictive recycling company, serving as an incentive to participate without payment. To sum up, the self-sustaining aspects are interwoven in both components (**Goal_{TG} 2**):

- By throwing away waste in the modified public trash can, new pictures are introduced in the system. The decision to use a public trash can was to minimize the privacy concerns reported in [299] and to reach a better scalability in terms of education (public trash can vs. trash cans at home).
- The user who wants to dispose of waste needs to decide how he or she would recycle it correctly by putting it into any of the bins, i.e., this represents a classification task to be done in front of the trash can.
- Users who are playing the mobile game receive the picture and need to classify it within the game.
- Both parties receive feedback (based on the majority decision of the players of the mobile game) and the result is further used in the respective game/gamification designs. Additionally, implicitly and consecutively (based on the findings of the study in Section 3.3.1), both parties should adapt towards this feedback, as they have seen whether or not they have

classified correctly and what would have been assessed as correct by the players. This is not only based on the previous results, but also on work [179, 184, 188] showing that in crowd-based scenarios, seeing results of others is a form of social influence and leads to adaption towards these results (because, for example, they think that others have better information). The rationale behind using such a setting is that not only people in front of the trash can, but also those who generate the feedback, might learn how to separate waste in a playful manner and will then do it more thoughtfully in their real life, even when not exposed to the system. As a side note: although it was shown in the study that the crowd error was small, it could happen that a wrong decision reached a majority. The mobile app especially has mechanisms that mitigate “wrong lessons” for the players (see below). Nonetheless, as an individual produces more errors than the crowd, as found in the study, providing crowd feedback in general seems better than providing no feedback.

Scenario: Before we introduce both components in more detail, the following scenario will give an example on the usage of the *Trash Game*:

Alice walks through campus and encounters one of the trash cans belonging to the Trash Game. She inspects the display and learns that her faculty has received many points during this week but is not in first place. She hopes that more people will sort their waste correctly to improve their faculty's score. Later, during a lecture, she receives a notification on her smartphone generated by the Trash Game mobile app. Her fictive recycling company, WasteGoneProperly, has received a new task and needs to handle a new object. Because Alice plans to improve her company with new upgrades, she needs to gain more in-game money, so she immediately reacts to the job, as this provides her with bonus cash. She knows that the trash cans produce these pictures and provide feedback based on the opinion of the crowd. She particularly likes that, as this helps people to learn how to do it correctly. Within the game she reacts to the new picture by stating how her company would handle this kind of product. Only then does she see how other players have classified the garbage. Unfortunately, the majority of the crowd decided differently. Alice wonders about this and later in the day she decides to dive into this topic. In the meantime, certain companies (i.e., other players) have received the task to provide evidence showing that their decision was indeed correct, and Alice is able to read their statements. Some arguments and references are accepted by the community, so she remembers how to classify the item and to do it correctly in the future.

Modified Trash Can

As introduced, unlike the *BinCam* approach [299] we focus on public trash cans. People in Germany are already familiar with trash cans consisting of multiple



Figure 3.13: An example public trash can for waste separation.

bins in public which also allow for a classification at the trash can itself. Normally, these bins are color coded, and printed pictures show representatives of the corresponding categories (see Figure 3.13 for an example). The design process of our trash can component was guided by the following requirements: the trash can needs to be able to recognize newly discarded objects; an image should be generated that only shows the new object (to simplify the classification task); the image should be made available online (in the mobile app); and the trash can should be able to display feedback in a playful way by using game elements.

Hardware design: We created a hardware prototype as a proof of concept. The basis is a wooden frame consisting of three different chambers and a smartphone attached inside the top of each (see Figure 3.14, left). Their purpose is to detect new objects and take pictures of them. A *Raspberry Pi* is used to do the data handling between server and smartphones and displays the results on the trash can's display. For persons using the trash can, the basic use stays the same – it is sufficient to throw the waste in one of the bins. The only difference to a normal trash can is that this insertion is recognized and visualized on a display.

Software design: The bin smartphones are always checking for differences in the RGB pixels taken by their cameras. If a difference is recognized, a picture is taken after two seconds (to allow time for the object to fall). Then, this picture is compared to the last picture taken to extract only the newly discarded object. This is done by identifying areas that have changed, and to reduce errors (as other content might also have changed position). We do a template matching and compare pixel arrangements of these areas with the pixel areas in the last picture, discarding similar areas (see Figure 3.14, right). Informal tests indicated that the algorithm is robust enough to produce pictures similar to those used in our study (see Section 3.3.1). These are sent to a server, which distributes them to the mobile apps. Our study has shown that even a small number of people can classify waste better than an individual. Thus, we consider the votes of the players after 30 seconds and if there are enough votes (20 is our current threshold)



Figure 3.14: Trash can prototype of the *Trash Game*: Left: Exterior/interior. Right: Image extraction algorithm from top to bottom: before insertions; after insertions; recognized differences; rectangle showing extracted picture.

we display this as final result on the trash can's display (otherwise it is shown as preliminary, with "*More time being needed to come to a consensus*"). The display visualizes the distribution of votes the group of players made (e.g., how many of the voters chose the categories). If a person does not want to wait until the time is up, he or she also has the chance to scan a barcode with a mobile device and is redirected to a web page showing the results there afterward. The system assumes that the category with the highest amount of votes is the correct one.

We create a competitive environment: each trash can is associated to a group, for example different faculties, divisions in a company or different supermarkets. After a correct (or incorrect) disposal (i.e., a (dis)agreement between crowd and chosen bin) the trash can receives (loses) points. The points are always clearly visible. In addition, each trash can shows potential CO₂ savings/production from the correctly/incorrectly separated waste inside it and the mobile game is

advertised as well. Different screens are shown if nothing is currently added: a leaderboard shows the points of the predecessor and the successor only for the current week (so that the leaderboard is not discouraging because the distances are too big [130]); the last additions are shown with the crowd voting; or a screen showing common disposal hints, and the advantages of waste separation being done correctly. All the screens are chosen to get people interested in the trash can itself and to provide them with gamified feedback, to get them engaged in waste separation and to potentially improve the learning outcome.

Mobile App

We designed the mobile app as a game with a purpose [314] (see also Section 2.4.2) to engage users (without paying them) in classifying pictures. The game is framed as a simulator in which the player is head of a recycling company that wants to become the market leader by quickly and correctly disposing of specific objects. The competition is amongst all players only (i.e., there is no Artificial Intelligence component). Occasionally, they receive tasks and the faster they react to them, the more money is generated, which can then be spent for improving the company. This improves their market value, but also has an impact on in-game mechanics (e.g., a higher market value means that players receive more tasks or gain more money per solved task). Conceptually, we distinguish three types of tasks, the *classification tasks*, *evidence tasks* and *knowledge tasks*. With different tasks we want to achieve a more varied game play but also improve the data quality. This was judged as important, as the previous study (see Section 3.3.1) also showed that the group decisions can be wrong. A player's profile in the background is used to save the times the player was correct and tracks demographic data, most importantly here the country, as this ensures that only pictures from trash cans located in the same country are provided to this player. The conceptual connection to real trash cans was also clarified in the application.

Classification tasks are triggered if a picture of a new object is taken by a trash can. Players receive a notification that a new classification task is available for them. By accepting it, they see the picture and the classification options (similar to the study in Section 3.3.1). In addition, they can provide a confidence value to their vote (see Figure 3.15). Only after they have selected an option, they can see how other players have decided, and they will be informed about the final decision for this object. If they belong to the majority, they receive money in a virtual currency. If not, it depends on how big the difference in the voting is. If it is small (indicating that the crowd was not completely sure), no money is lost; otherwise, because "*The company needs to recycle the object in a more expensive way*", some money is subtracted. The reason for showing the crowd distribution lies in the fact that people who see the distribution might be motivated to engage (e.g., if they see that the vote was only slightly in favor of the other option) either in the courses or in the *evidence task* (see below). We used the strict expert aggregation, as this performed best in our study (see Section 3.3.1).



Figure 3.15: Classification screen of the app after a decision (landscape-view).

To support a deeper engagement with recycling, we integrated *evidence tasks*. Players are selected for them randomly, but can also chose to do these voluntarily. The main goal is to provide evidence (e.g., official links and/or reasonable explanations) for a previously classified object that either supports the crowd opinion or contradicts it. Similar to the *Stack Exchange network*³¹, all players who participated in the crowd voting can see this discussion and up- and down-vote the different contributions, and finally accept an answer. Involved players receive a notification and potentially a (virtual) refund. In addition, the company which provided the accepted evidence receives a reward that positively affects their market value. This task was created to account for crowd errors, which we also saw in our online survey, especially for uncommon objects. The third task type are *knowledge tasks* in which questions are provided and players either contribute answers (e.g., “What is this object called?” or “Do these pictures show the same objects?”) or assess already-given answers (e.g., “Are these valid names for this object?”), which improves a knowledge base in the background. This can later serve to change the game mechanics and to inform machine learning algorithms. Besides these tasks, players have the option to attend *courses* in which common wrong classifications are presented with explanations. Attending a course provide players a virtual certificate and virtual bonus money, if they classifies such an object correctly in the future. To keep people engaged, we ensure that specific pictures that were already classified by the crowd (but not by the player) are given to the players periodically even if nothing has been recently discarded.

³¹ *Stack Overflow* as described in Section 1.2.1 is part of this network.

3.3.3 Concept Evaluation of the *Trash Game*

To evaluate the *Trash Game* in all its aspects we would have needed to receive access to a competitive environment where we could place, for example, several prototypes of the modified trash can, and would have needed to distribute the mobile game over app stores to gain a significant player base. As the current proof of concept of the trash can is not robust against thievery (as it would be easy to remove the smartphones or the *Raspberry Pi*) and putting an app into a store does not guarantee many players instantly (which would be necessary to evaluate the self-sustaining system “in the wild”), we decided on a concept evaluation first to learn how the prototypes are perceived.

Method

We presented both prototypes, the trash can and the mobile application, to university employees, students and visitors at the cafeteria foyer around lunchtime. They experienced the process by throwing an object (we provided several) into one of the bins and could also vote with the mobile app. The crowd feedback was provided in a *Wizard of Oz* style experiment [54] in which we select which feedback is shown (potentially adjusted by the participant vote). The selection was based on the results from our online study (see Section 3.3.1), e.g. if a supermarket receipt was inserted, we presented the crowd classifications for this object. This means we used real crowd values. While showing the process, we also explained the concepts of the prototypes and answered questions. Subsequently, the participants were provided with a questionnaire, consisting of free text questions and questions to be answered on a 7-point scale, with the labels *strongly disagree* and *strongly agree* shown on the extreme values.

Results

35 people participated in our evaluation (12 female, 23 male; age: <21: 16, 21–30: 13, 31–40: 4, 41–50: 1, >50: 1). Questions concerning their waste separation behavior were answered similarly to our online study. Participants tended to agree to that waste separation is easy ($M=5.1$, $SD=1.2$, $Mdn=5$), that they do it correctly ($M=5.2$, $SD=.9$, $Mdn=5$) and to the best of their knowledge ($M=5.6$, $SD=1.5$, $Mdn=6$). They had mixed feelings about whether waste separation in Germany is complicated ($M=3.8$, $SD=1.9$, $Mdn=4$) and they do not seek more information if unsure how to separate waste correctly ($M=2.7$, $SD=1.7$, $Mdn=2$). Concerning the game concept, participants liked the trash can ($M=5.8$, $SD=1.6$, $Mdn=6$) and the mobile app ($M=5.7$, $SD=1.6$, $Mdn=6$). If they had the chance to decide to dispose of their waste in a normal or the augmented trash can, they would use our prototype to receive feedback ($M=5.4$, $SD=1.7$, $Mdn=6$). Eleven participants said this was because they liked the idea and four stated that it helps the environment. Three participants stated that the waiting time is a problem.

We elaborated on this further and learned that eight would wait, 14 would wait if feedback is provided quickly, six would not wait at all and two reported that they would wait only if unsure whether they have disposed of correctly. Only one participant reported that he or she would not consider the mobile web page as an alternative if the waiting time is too long. Here, a solution could be to integrate approaches similar to [19], in which a crowd constantly is activated to solve tasks faster, but in our case with only virtual incentives. In general, the crowd feedback is judged useful ($M=5.6$, $SD=1.3$, $Mdn=6$) even though it could be wrong (which was also demonstrated). On the other hand, the responses were mixed for the question whether participants would also let the crowd classify their waste at home ($M=3.7$, $SD=1.9$, $Mdn=4$), mostly because of privacy issues, as the free-text answers showed (13×). This is a replication of the results of the *BinCam* [299]. Participants wished for further graphical elements, a variable design for different age groups, a mechanical component in which the waste is stored until the crowd has decided, and a way to achieve bonus points. They are undecided whether to play the mobile game ($M=3.9$, $SD=1.9$, $Mdn=4$). A reason for this could be that only a part of the app was presented and the other functions were only explained. The answers showed that people focused more on the classification part inside the app and did not consider that these tasks were integrated into the game play. In contrast, they liked that they could have an influence on the feedback the trash can shows ($M=5.5$, $SD=1.6$, $Mdn=6$). Concerning the functions, participants suggested a social media connection to compete with friends and to integrate challenges, and a collaboration with real recycling companies as experts providing a more reliable ground truth.

Discussion

Both the trash can and the mobile application concepts were perceived positively in this concept evaluation (**Goal_{TG} 2**). An especially interesting result is that even though we demonstrated that the crowd can be wrong, participants still judged the feedback as useful. Given the nature of the self-sustaining system (and the results of our online study), we would have assessed this as the most severe issue in the design. Even though the mobile app has concepts that might correct such wrong classifications eventually, from an immediate perspective this might lead to frustration for participants that were sure of being correct. Furthermore, the local crowd (i.e., people that use the trash can) might not profit from the correction anymore, as the system has no trace of who actually has disposed of something. Nonetheless, it seems that having feedback at all is perceived better than not receiving (potentially erroneous) feedback. Considering the results of our online study, we also see that the error rate for crowds is comparatively low and in combination with this result, we see this as acceptable. It also needs to be kept in mind that contradictory information (my knowledge vs. what the *Trash Game* says) might also lead to more engagement (as for example Weiner [322] describes effects on people's motivation depending on what they attribute their

performance to). Eventually this might lead to a check on the waste disposal rules to find out who is correct. From a limitations point of view, the major limitation that needs to be kept in mind is that we only conducted a concept evaluation in a *Wizard of Oz* style experiment [54], i.e., the interaction was restricted. In addition, our sample was younger and thus it is not clear how an older population would assess the concept.

3.3.4 Contribution to the Thesis' Questions

Although not tested “in the wild”, the concept evaluation of the *Trash Game* provided encouraging results. Together with the results of the conducted online study, it shows that the idea of a self-sustaining system to empower human recycling capabilities is reasonable. Individuals participating in the classification tasks improve implicitly over time and with the presented system design the waste separation capabilities of people at the trash can and of the players playing the mobile game can be likewise targeted (**Goal_{TG} 1**). If such a system were rolled-out on a large scale, many people could be reached, which in consequence would have positive effects on the environment. But the self-sustaining system design provided further beneficial aspects beyond the education aspect itself (**Goal_{TG} 2**). The crowd provides more reliable information on the contents disposed of in the bins; more specifically, through the knowledge tasks, we might also achieve a complete classification of waste inside the bin, which could enable recycling companies to better decide what should be done with the contents themselves. Moreover, with such an approach it could also be possible to learn from errors and share this knowledge with policymakers to help them to improve their own (educational) publications. In addition, it could be of interest for companies to deploy information within our app to improve crowd performance.

The result shows that a self-sustaining system also has an influence on its users (adding to **RQ1**; see Section 1.4): users not only impact how the system behaves, but the system also impacts how they behave (underlining the reciprocity aspect). Here, this is exemplified with the learning effect that we saw in the online study. Although we never told the participants that they would be re-tested, by just seeing what would have been correct, they adapted. Similar results can be expected from the proposed system design, as neither the trash can nor the mobile app state that they are educational in nature or require participants to learn. Besides the personal knowledge gain in how to recycle correctly, it needs to be emphasized that this further improves the self-sustaining system. Users become better at the task over time, strengthening the core idea of users being an important component of the system from which they receive a service. An interesting aspect that was revealed in the concept evaluation is that many participants reported liking to have an impact on the system, i.e., that their opinions impact which bin is actually shown as the correct one for recycling the object. This underlines the core ideas this thesis follows: users want to have more influence in a system, in this case even on the fabric of the service itself.

3.4 Summary

In this chapter, we have analyzed gamified self-sustaining systems, i.e., systems following crowdsourcing concepts that only provide their service through integrating their users as an important component into their system design. With *ExpenseControl* we validated the idea of a self-sustaining system that utilized microtasks that were not engaging and that did not give users any feedback about their solutions. Even in such a setting and without additional motivational mechanisms, we saw that people engaged in the solving of microtasks, although this was never explicitly formulated as a study task. Through the user study of *ExpenseControl*, we found that participants already solved a fair amount of microtasks, even when not receiving an additional layer of motivation. Nonetheless, the usage of gamification in this setup motivated the participants further and doubled the average amount of solved tasks. Using game-based motivation thus seems a reasonable idea in self-sustaining systems as well. With the *Trash Game*, we focused on the question whether users in a self-sustaining system also gain further benefits by working on the tasks. Through a user study we showed that people, even if never told that we want to educate them, improved at the task of classifying pictures of waste. Feedback from the self-sustaining system is thus helpful, as it improves user input. In consequence, this also improves the system itself, as subsequent inputs are of higher quality.

Overall, this chapter considered systems in which users have a strong influence, adding to **RQ1** (see Section 1.4). Although a group effort, the connection between individual users can only be considered as weak. The existence of other users is known (e.g., through the leaderboard in *ExpenseControl* or the distribution of votes in the *Trash Game*), but it is of no real consequence for users. They do not need to interact with each other and have no immediate influence on one another. The users also had no influence on the motivational aspects of the system (e.g., which kind of gamification elements should be used), nor a direct influence on how the system achieves to its outcome (e.g., how specific receipt entries are corrected by *ExpenseControl* or which classification is assumed to be correct by the *Trash Game*). Considering the first aspect, in *ExpenseControl* we already saw that the fixed gamification approach did not motivate everyone similarly (see also Section 2.3). What happens if individuals are empowered to influence the game-based approach at the runtime of a system is considered in the next chapter. For both systems, we have shown that the overall performance depends on the aggregation mechanisms used for combining individual inputs (being in line with the approaches reported in Section 2.4.3). For self-sustaining systems, it is obvious to use the algorithm that provides the best possible outcome. This can easily be decided when the system goals are clear. In contrast, in Chapter 6 we consider settings in which the goals are manifold, and thus how to aggregate is not obvious. In this chapter, we investigate the influence that a group can have on individuals and how groups can self-administrate themselves in such systems, e.g., by adapting the aggregation mechanisms.

Chapter 4

Self-Tailored Gamification

This chapter introduces the idea of “bottom-up” gamification. Here, users are given the option to customize the gamification in a system as they see fit. In contrast to the customization options presented in Section 2.3.2, “bottom-up” gamification offers more fundamental options, as users can for example decide to not use gamification at all, or to completely change the game design approaches during their system usage at runtime. “Bottom-up” gamification fits with the thesis scope and our research question **RQ2** (see Section 1.4), as users have a significant influence on the motivational approach chosen in a system. After defining “bottom-up” gamification and discussing which game elements are suitable for it, we present several user studies. These not only provided requirements but also revealed positive qualitatively and quantitatively measured effects for self-tailored gamification. Thus, they indicate that it is beneficial to provide users with options to adapt gamification settings on their own.

Section 4.2 and Section 4.3 are based on the publication [169], Section 4.4 is based on [170] and Section 4.5 is based on [167].

4.1 Introduction

As illustrated in Section 2.3, an ongoing effort in current gamification research is to move away from “one-size-fits-all” gamification systems, in which every user is presented with the same set of game elements to a tailored experience, in which every user has an individualized set of elements. As also discussed there, such approaches are shown to increase the effects gamification induces. Personalization (the system adapts automatically or is adapted by researchers; see Section 2.3.1) and customization (the system can be adapted by the user; see Section 2.3.2) are two options to tailor gamification. As elaborated on in the

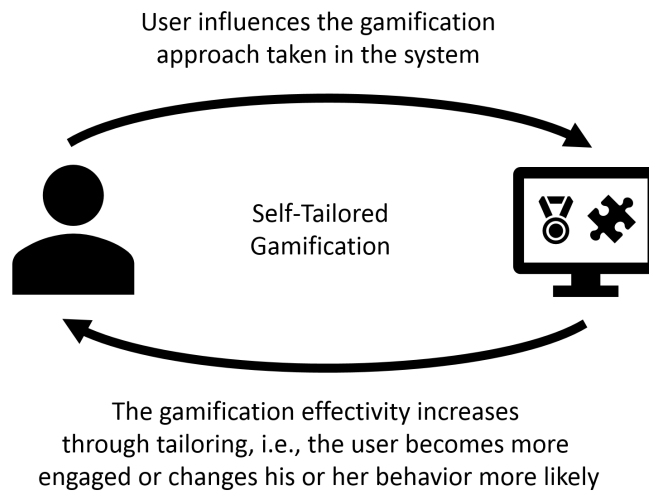


Figure 4.1: Instantiated schematic of reciprocity in self-tailored gamification.

corresponding related work sections, while personalization means less effort for users than customization (but also provides less autonomy) it also needs to account for several aspects in relation to individual and contextual factors. So far, to our knowledge, no personalization solution exists that covers all these aspects within one system.

In Section 2.3.2, we highlighted on why customization is an alternative (such as making an activity meaningful for a user) and on approaches that allow its users to adapt parts of the game elements. In this chapter, we will go one step further and consider an approach in which every user receives fundamental customization options. Besides allowing users to decide at runtime if they want to use gamification at all, we offer full customization options: users cannot only adapt game elements but can also decide which elements are available in the system. Thus, they can combine and adjust elements as they see fit. Considering the scope of this thesis, enabling users to impact the gamification approach a system offers at its runtime provides users significant influence options. Independent of the exact goal for which gamification was implemented (e.g., raising the engagement to use the app or to achieve a behavior change), users can now alter it to their needs and thus are potentially more motivated and likely to reach the goals. This directly connects to the autonomy aspect of the *Self-Determination Theory* (SDT) (see Section 2.2.2). Thus, in consequence, through empowering users to influence this system aspect, they have an influence on themselves (**RQ2**; see Section 1.4). Figure 4.1 shows an instantiated schematic of this reciprocity in self-tailored gamification. In contrast to the self-sustaining systems in Chapter 3, in which we considered loosely coupled groups, and Chapter 5 and Chapter 6, where tightly coupled groups are considered, customizing gamification is an individual aspect. Although some game elements can only be reasonably used with other users in a system (e.g., competition), every user defines *his or her* gamification for *him- or herself*, i.e., input of one user has no direct impact on others.

We termed this concept “bottom-up” gamification. We define it in the next section and we discuss common game elements in terms of whether they are suitable for such an approach. This is followed by presenting a study that evaluated the “bottom-up” gamification idea a priori (without using a prototype) and provided first requirements towards “bottom-up” gamification. Based on these results, we created a “bottom-up” task management application (*BU-ToDo*), evaluated it and found that “bottom-up” gamification is perceived positively and, based on self-reported data, subjectively leads to adaptations of the participants’ behavior. Additionally, to evaluate the effects quantitatively, we utilized a microtask setting similar to the one used in *ExpenseControl* (see Section 3.2) in an online platform (*BU-Microtasks Platform*). Here, we were able to show that “bottom-up” aspects indeed have positive effects, as participants who used their available choices solved more microtasks, in contrast to using no gamification or a fixed gamification setting. As prototypes always come with certain limitations (which is necessary from an implementation point of view) to the actual choices users have (i.e., a finite set of game elements to select from), we conclude this chapter with a study in which users were not restricted. With it, we investigate what kinds of gamification settings are created by users, and start a discussion around which aspects account for the positive effects of “bottom-up” gamification.

4.2 “Bottom-Up” Gamification

We define “bottom-up” gamification as follows:

“Bottom-up” gamification allows users to decide whether they want to use gamification at a system’s runtime. If they do, they can adapt all available game elements in the system and combine them as they see fit.

The definition highlights the adaptability of game elements in systems at their runtime. The opposite of “bottom-up” gamification is “top-down” gamification, which denotes gamification approaches in which people who are typically not the later users of the system (e.g., superiors or game designers), design and orchestrate the game elements. As discussed (see Section 1.3), “top-down” gamification in the work context has led to the term “*exploitationware*”, a criticism that might not arise in settings in which users can tailor systems to their own needs. “Top-down” approaches might also offer personalization or customization options, but concerning the latter, the adaptation degree is usually restricted (see Section 2.3) and users cannot turn off gamification in such systems.

“Bottom-up” gamification has some overlap, but should not be confounded with, “*bottom-up game design*”³². This concept focuses on the programmers and describes a game design process in which mechanisms are implemented first (e.g., how

³² Gamasutra: *The Designer’s Notebook: The Perils of Bottom-up Game Design*,
<https://goo.gl/XdeMXh> (last accessed: 2018-07-07)

to simulate an escalator scenario in a building) and only after, it is considered how to make this scenario fun or game-like. In our case, a user should not have to implement something (but should receive an easy way to access different game elements), but will also consider how to make an activity fun by selecting suitable game elements. From a design point of view, “bottom-up” gamification is more than *user-centered design* (see Section 2.2.5) in gamification [219]. Here, normally, users and experts are involved; while the users provide data, experts interpret these to understand how a problem needs to be solved. In our case, while *user-centered design* might be still involved (e.g., to inform which game elements might be offered in such a setting in general), overall, users can solve their problems on their own without any mediation by experts. This is a form of “*user-led design*”³³, but in our case, one that happens at the runtime of the system.

4.2.1 “Bottom-Up” Game Elements

Although the definition above suggests that users can decide which game elements they want to use, it seems unfeasible to offer all possible game elements in a system from a usability (e.g., because of information overload [205, 304]) and a choice overload perspective (see Section 2.2.3). Additionally, it appears that not all gamification elements are equally suitable for “bottom-up” gamification as, for example, the adaptations would be too demanding, or adapting them would break the core aspect of a game element (e.g., the element *surprise*). We considered the literature overviews in [106, 134, 269] and analyzed the presented commonly used major game elements for their suitability in a “bottom-up” approach, i.e., whether users can adjust the element to their needs and whether the element is then still reasonable to use:

- Receiving *points* seems suitable for a “bottom-up” approach, as users could easily assign points for solving tasks. This element was already used in a “bottom-up”-like approach [136] in which users set up and receive experience points for solving tasks in a task management application. An issue that might appear is that the amount of points, when user-generated, is not easily comparable across users.
- *Achievements*, such as visual badges and/or textual ranks, could also be added as a reward after solving a task. The (self-)creation of visually attractive badges in a “bottom-up” process is theoretically possible (and can be supported by a badge creation tool). Nevertheless, badge placement is an important aspect here, i.e., finding a balance between task difficulty (neither too easy nor too hard) and receiving the badge [7]. As users can decide what they can/cannot achieve, this seems feasible. The positive impact of achievements in general was shown in [103] and [194].

³³ UX Magazine: *User-Led Does Not Equal User-Centered*,
<https://goo.gl/BvBnQ8> (last accessed: 2018-07-07)

- *Self-defined rewards* are rewards that are not only available virtually and are defined by a user (e.g., buying a new CD after finishing a task). This is an aspect that can easily be integrated in “bottom-up” systems. In general, these motivate users extrinsically [16].
- The elements *progression*, *clear goals*, and *feedback* in the form of receiving a reward for advancing in a task, having the option to set up rewards that are tiered, and seeing how much progress was made towards a specific goal can be added into a “bottom-up” approach. The application needs to assist users in setting up these elements, and users should be able to define the specific rules to reach a set goal, e.g., how many points are necessary to reach a new level, or when they want to unlock intermediate rewards.
- *Competition*, *cooperation*, and *social recognition* are another group of elements that are also possible in a “bottom-up” approach. Competition, according to [119, 220], can be a motivator or demotivator depending on the player type [306]. It is important for a “bottom-up” approach that the participation in competitions remains voluntary, as peer pressure might be an issue otherwise [102]. As competitive challenges involve other users, there is a need for some kind of anti-cheating mechanism. For specific tasks, this can be done to with sensory input (e.g., “Who is running more in a week?”). For other tasks, making users submit proofs which are reviewed by another person or an automated system before rewards are received is also an option in a “bottom-up” scenario. The competitive element should also be seen in conjunction with *leaderboards* in which performance is directly compared to others [130], or used implicitly, e.g., by showing how often a badge was collected by other users. As long as leaderboards are task-specific (and thus rewards are the same for everyone), they also remain comparable. Cooperation could be implemented by offering tasks that can be solved together with other users. Progress could also be checked by the same mechanisms as mentioned before. *Social recognition* can also be motivational, especially as evidence exists that virtual rewards such as points become more meaningful when shared [105]. In a “bottom-up” sense this could be realized by (voluntarily) informing people that a task was started/finished or that rewards were received.
- *Story elements* range from simple descriptions of situations [102] to the simulation of complex worlds in which the user can improve a *virtual avatar* through rewards or *virtual goods*. The avatars could then also be used to, for example, compete against other users’ avatars. For a “bottom-up” approach, story elements do not seem easily usable, as these impose more work on the user. Additionally, it appears unreasonable for the same reason that the element *surprise* seems unsuitable: users could be allowed to create (story/surprise) content and to make the results accessible to others. In comparison to other elements, though, it seems difficult to justify why users should generate something that is demanding to create and cannot be

expected to add much to the user's motivation when unlocked (as the story part or the surprise is already known to the user). Using an avatar and virtual goods without story elements is possible, but then appears similar to achievements (i.e., you have achieved the unlocking of a specific virtual good which has the same effect as receiving a badge). Self-expression through the creation of virtual goods could be beneficial for motivation, but also imposes more effort for the user in the application, as opposed to mere visual badges. Customizing avatars (e.g., changing the appearance of them) was shown to raise the users' self-identification with them and relates to positive effects later on [21]. While being able to customize an avatar seems to be suitable for a "bottom-up" scenario, it raises the question of whether users should also be able to add new assets (e.g., new clothing) or customizable dimensions (e.g., attributes), which again, might impose more effort compared to other game elements.

4.2.2 A Priori User Assessment of "Bottom-Up" Gamification

To learn about the perception of "bottom-up" gamification, we decided to conduct a user study with the following goal:

Goal_{BUExpectations} *Investigation of "bottom-up" gamification without using a specific prototype:* By validating the "bottom-up" idea by not using a prototypical realization, participants can assess the idea without being biased by a concrete realization [227], i.e., we can assess expectations. This, on the other hand, provides an overview of whether the idea of self-tailored gamification is generally reasonable.

Following this goal, we wanted to gain insights for the following questions:

- Q1** Are people open to using (gamified) applications for motivational purposes?
- Q2** Can people imagine defining their own gamification?
- Q3** Are there differences in how people gamify on their own?
- Q4** What are requirements for "bottom-up" applications?

Q1 should provide insights on whether users would be willing to use applications to motivate themselves. As we will provide digital systems that offer "bottom-up" gamification, this is an important prerequisite. **Q2** and **Q3** will validate the idea of "bottom-up" gamification and whether people behave differently when they can gamify on their own. Given positive answers to the previous questions, **Q4** will assist in the development of applications that provide "bottom-up" approaches.

Scenario	Abbr.	Explanation
Cleaning	<i>Cl</i>	Imagine the kitchen is due for cleaning. You want to motivate yourself with the app to get this done.
Piece	<i>Pi</i>	Imagine you work for a manufacturer and build furniture by piece work. The work is monotonous and you want to use the app to make it more exciting today.
Exercise	<i>Ex</i>	Imagine you want to exercise more, i.e., you want to go for a run multiple times a week. You want to use the app to motivate yourself to reach this goal.
Energy	<i>En</i>	Imagine you want to save energy at your company, e.g., by turning off the lights after work. You want to use the app to motivate yourself to reach this goal.

Table 4.1: Scenarios used in the study, their abbreviations and explanations.

Method

We set up an online questionnaire in German with questions covering demographics, participants’ gaming affinity, their experience with gamification in general, how they motivate themselves day-to-day to do unpleasant tasks and how they perceive “bottom-up” gamification in different contexts. To this end, we presented the following abstract idea of a mobile application using “bottom-up” gamification (and this explanation of what game elements are, i.e., *components describing basic mechanics often implemented in games, such as points, badges, leaderboards, levels, avatars, story elements etc.*):

Let’s assume for the following questions that there is an app which allows you to execute every task of your life (including in the workplace) in a playful manner. The app allows you to manually define tasks (e.g., doing laundry) or to accept already existing tasks in order to accomplish them. In order to make tedious tasks more exciting, you are able to do tasks in a playful manner and you can also decide on the shape and form this will take. Depending on your personal demands, you may establish game elements and rewards as you wish (e.g., points and a leaderboard), while solving tasks either alone, in cooperation or competing against others.

We then provided four scenarios (see Table 4.1) in which participants were asked to indicate whether they could imagine motivating themselves in a playful fashion in every scenario. If they could, they were presented with twelve different game elements and were to state whether they would want to use these in those situations. Scenarios were selected to cover work and private life, as well as one-time tasks, i.e., tasks that need to be done but most likely will not affect the person further (*Cl*, *Pi*) and behavior change tasks, i.e., tasks that need be done more often and might lead to a behavioral change (*Ex*, *En*). We provided scenario-specific framing for the game elements, to explain these further for the

participants. For example, the framing for *Competition* in *Cl* was “*Challenge other app users to clean their kitchen, the most effective/fastest wins*”, while in *En* it was “*Challenge other users to also save power. The user with the highest savings wins*”. The questionnaire consisted mainly of questions to be answered on a 5-point scale with labels shown on every option (*disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree*), yes/no questions and (optional) free text questions. The questionnaire was promoted via mailing lists and social media channels.

Results

75 participants (46 male, 29 female; age: <21: 2; 21–30: 55, 31–40: 9, >40: 9) completed the questionnaire. Mostly participants reported being students (48%) or employees (37%). 69 (92%) own a smartphone or tablet and assess themselves as experienced with it on a 5-point scale ($M=4.4$, $SD=.7$, $Mdn=5$). Concerning their affinity to games, 43 (57%) reported playing video games regularly, for an average of 8.9 hours per week ($SD=8.4$), and 40 (53%) reported playing parlor games regularly (2.7 hours per week, $SD=3.3$). Only 19 participants (25%) already knew about gamification, they tended to have positive experiences with it ($M=3.6$, $SD=1$, $Mdn=4$), and tended to agree that it motivated them to use systems which offered it more ($M=3.6$, $SD=1$, $Mdn=3.5$). Before the “bottom-up” idea was introduced, we asked the participants what applications that aim to provide motivation for tasks should or should not offer (to be answered in free text fields). 55 different aspects were mentioned, but we only considered those mentioned at least twice. Table 4.2 shows the elicited requirements (R1–R11). R12–15 represent requirements we have derived through the analysis of the results made in the questionnaire. These requirements add to **Q4**.

Tasks today: 62 participants (83%) reported having unpleasant tasks in their private and 45 (60%) in their working life. Some participants additionally reported how they motivate themselves to do them: the knowledge that the task simply needs to be done (18× mentioned in the free text field), establishment of a motivational atmosphere (9×), using a checklist to see progress (7×), rewarding themselves with something after the task (6×), prospective joy about solving the task (6×), social pressure (5×) and receiving an external reward (grades/salary) (5×). An application that motivates them to complete tasks in their private (answered by all participants) or working life (answered by 57) could be imagined by 46 (61%)/39 (68%). We also asked all participants whether they could imagine solving tasks in a playful fashion in their private or working life to which they tended to agree to ($M=3.7$, $SD=1$, $Mdn=4$ / $M=3.5$, $SD=1.4$, $Mdn=4$). Participants who play games regularly did not answer these questions significantly different.

Based on these results, offering sources of motivation seems beneficial. As (more than) half of the users could imagine using an application for motivational purposes, gamified assistance could be such a source (see **Q1**). How participants motivate themselves varies; thus, a flexible approach that is able to address various needs seems beneficial as well (supporting requirement **R8**; see Table 4.2).

ID	Requirement
R1	The app should offer a to-do-list-like interface (10x) and provide an overview for already-done tasks (4x).
R2	The app should not generate overhead, i.e., adding/handling tasks should not take longer than doing the task (12x). This requirement was also supported by the other results gained by the questionnaire.
R3	The app should offer the option to formulate intermediate goals and achieved progress should be visualized (4x).
R4	The app should offer social elements like collaboration, competition or the knowledge that other users can see a user’s own progress (6x).
R5	The app should offer a reminder functionality (3x).
R6	The app should offer a timer functionality, which could also be used to improve one’s own performance over time (5x).
R7	The app should offer rewards for tasks (7x), such as achievements (3x), points (5x), or real incentives (3x) and unlocking them should be possible (2x).
R8	Users should be able to decide which functionality they want to use (2x) and the app should offer enough options to provide flexibility and variance (2x). This requirement was also supported by the other results gained by the questionnaire.
R9	The app should not put pressure on users (3x) or dictate when tasks are to be done (6x).
R10	The app should allow users to customize the frequency of notifications (10x).
R11	The app should allow users to share only data they want to share (12x).
R12	The app should be usable in a domain-independent way, as the results show that users want to gamify different parts of their life.
R13	The app should make it easy to use game elements and if elements exist that request user-generated content (such as stories), it should be easy for users to add new content.
R14	The app should offer the option to let task completions be reviewed by others. For a “bottom-up” approach, a user should always be able to select whether a reviewer should be integrated.
R15	The app should offer the option to inspect how other tasks have been gamified (without contradicting R11), as inspiration for users that might not know how to gamify them.

Table 4.2: The requirements derived from the online study. R1 to R11 were mentioned by participants in free text fields (numbers in parentheses denote how often) before we introduce the “bottom-up” idea, while R12 to R15 were derived by us based on the participants’ answers to the closed questions.

Perception of “bottom-up” gamification: Participants tended to assess the idea of “bottom-up” gamification positively ($M=3.5$, $SD=1.2$, $Mdn=4$). Chi-squared tests showed that people who claim to be open to using an app for motivation in their private/work life were significantly more open to our app concept ($\chi^2(4, N=75)=32.2$, $p<.001$, $V=.65$ / $\chi^2(4, N=57)=28.1$, $p<.001$, $V=.7$). The same is true for people who want to solve tasks playfully in their private/work life ($\chi^2(16, N=75)=90.3$, $p<.001$, $V=.55$ / $\chi^2(16, N=75)=34.5$, $p<.05$, $V=.34$). This indicates that our “bottom-up” approach fits into users’ expectations (see **Q2**, **Q3**). We also asked whether participants want to use a motivational app in all areas of their daily life, and received mixed answers with a favorable tendency toward it ($M=3.5$, $SD=1$, $Mdn=4$). Nonetheless, this highlights that a domain-independent approach, that focused not just on one setting, seems beneficial (establishes **R12**).

The 69 participants (92%) who did not completely disagree to using such an app were presented with further questions. Concerning the game elements (see **Q2**), the wish for influence was expressed, as our sample tended to want to select game elements on their own ($M=3.9$, $SD=1$, $Mdn=4$; supporting **R8**). The sample was indifferent on whether they would create new content (e.g., *story elements*) ($M=2.8$, $SD=1.2$, $Mdn=3$) and indifferent on whether they would want to think more about the game elements ($M=3.2$, $SD=1.2$, $Mdn=3$; supporting **R2**). This indicates that a “bottom-up” system should offer a rich variety of game elements and should not necessarily integrate elements that are too demanding in terms of user-generated content (which establishes **R13**). As in a “bottom-up” approach, it is questionable how to handle rewards in general, we asked whether the participants wanted to decide for themselves or let others decide on a reward, but the answers were inconclusive ($M=3.2$, $SD=1$, $Mdn=3$ / $M=3.1$, $SD=1$, $Mdn=3$). We also integrated questions on whether there should be anti-cheating mechanisms (as a multi-select question): 6% think that monitoring whether or not a task was fulfilled correctly (before receiving the reward) is necessary for tasks in which only the user is involved, 20% think that monitoring is unnecessary, 36% stated that it is necessary whenever other users are involved (e.g., in competitions) and 38% think that a check is always necessary. Some kind of review mechanism that can be activated by the user, if requested, seems reasonable to comply with these views (establishing **R14**).

Scenarios and game elements: We analyzed how participants perceived the scenarios (see Table 4.1). *Cl* was not perceived as convenient ($M=2.6$, $SD=1.2$, $Mdn=3$), nor was *Pi* ($M=2.6$, $SD=1.2$, $Mdn=3$). The other two scenarios were perceived as significantly more convenient (*Ex*: $M=3.5$, $SD=1.3$, $Mdn=3$, *En*: $M=3.5$, $SD=1.1$, $Mdn=3$), as a Friedmann test with step-down follow-up analysis revealed for these two groups ($\chi^2_F(3)=32.5$, $p<.001$). For every scenario, participants were asked to indicate whether they could imagine completing the corresponding task playfully (as a yes/no question). 62 participants (83%) could do so in at least one scenario. If they disagreed, we asked for reasons. For scenario *Cl/Pi/Ex/En*, the 28/37/21/53 participants that disagreed (18%/11%/38%/43%) stated that the task itself is motivating for them and 57%/63%/33%/48% stated that they cannot imagine a game in this scenario. While the latter indicates that hints on how specific tasks could be gamified might be helpful (establishing **R15**), the first indicates that a domain-independent “bottom-up” approach, in which users can decide for themselves *whether or not* they want to gamify an experience, seems beneficial (supporting **R12**).

For every scenario, participants who could imagine a playful approach for it were asked to select game elements they would use (see **Q3**). We checked how much individuals vary per scenario by averaging over all ratings per element and participant. Table 4.3 shows that they do not vary at all ($SD=0$), or only slightly ($SD<1.0$). This suggests, at least for our scenarios, that most participants would stick to the same game elements for motivating themselves. But the table also shows which elements would be selected over all participants in general. From

	Game element variation		Acceptance rate per scenario			
	$SD_{=0.0}$	$SD_{\leq 1.0}$	<i>Cl</i>	<i>Pi</i>	<i>Ex</i>	<i>En</i>
Receiving points	50%	85%	74%	68%	70%	64%
Virtual character receiving a benefit	52%	94%	64%	66%	61%	50%
Receiving badges	59%	88%	72%	66%	70%	69%
Unlocking new functions inside the app	40%	80%	60%	53%	61%	50%
There is a narrative setting around the task	33%	80%	49%	53%	50%	59%
Seeing a progress bar	46%	85%	87%	82%	89%	68%
Competition against other users	22%	69%	57%	74%	59%	73%
Cooperation with other users to do tasks	24%	76%	51%	82%	67%	73%
Receiving a reward defined by myself	43%	83%	53%	29%	41%	18%
Receiving a reward from friends/employer	43%	70%	53%	82%	37%	73%
Informing friends about starting the task	65%	83%	10%	18%	17%	9%
Informing friends about finishing the task	52%	83%	17%	5%	19%	27%

Table 4.3: Game element variation across all scenarios and participants that selected game elements (asked on a 5-point scale) for at least two scenarios and acceptance rates per scenario (% of participants responding with 4 or 5 to the question of whether they would use the element here).

this, we can conclude that nearly every element we asked for can be of relevance across participants in specific scenarios, which also supports **R8** and **R12**. This is in line with the reported amount of individual and contextual differences in the perception of game elements (see Section 2.3). Interestingly, competition, cooperation and receiving rewards (from myself/from others) seem more likely to be influenced by the context. Additionally, the elements that inform others about starting/finishing tasks are not perceived well by the participants in general.

Discussion

Concerning **Goal_{BUExpectations}**, this study did not use a specific prototype to validate the idea of “bottom-up” gamification, in order to assess expectations without introducing a bias through the system “hosting” it [227]. The study results provide evidence that the concept of “bottom-up” gamification is reasonable to follow. It was revealed not only that participants motivated themselves to do unpleasant tasks through different means, but also that more than half of the participants could imagine solving tasks in a playful fashion. Some participants reported already using game elements for motivation, such as seeing their own progress on checklists, or rewarding themselves. Overall, the sample was open to defining their own gamified experiences to motivate themselves. Considering the literature on the *SDT*, the role of autonomy and the positive impact of choice (see Section 2.2) as well as the positive reports of other customization approaches (see Section 2.3.2), this was expected. Tailoring also seems necessary, as we learned across different questions that there is a need for defining or selecting one’s own

game elements to account for individual differences. Interestingly, individuals seem not to vary much in the scenarios in terms of which game elements they want to use, i.e., a participant often selected the same elements across scenarios. At the same time, we learned that every offered game element was assessed as interesting to a certain extent, hinting that many different elements might be necessary in a “bottom-up” system to appeal for a broad user base. Finally, the study also contributed requirements for “bottom-up” applications. Concerning limitations, we had only German participants, which is a threat to external validity. Second, the sample consisted mainly of younger gaming-affine people. Thus, it is questionable how these results transfer to an older population with potentially less affinity to games, but as the desire to play is age-independent (see Section 1.1), this seems acceptable.

Overall, based on these encouraging results, we developed and tested an application offering “bottom-up” gamification (see below). We will discuss the implications of the online study towards the thesis’ questions together with the findings made with this application in Section 4.3.3.

4.3 A “Bottom-Up” Task Management Application

As the previous study only assessed expectations and validated the acceptance of “bottom-up” gamification in general, we were not able to state what effects the approach would have on users themselves. Consequently, to validate the concept further, the following goals were targeted:

- Goal_{BUToDo} 1** *Creation of an application offering “bottom-up” gamification:* By creating an application making use of “bottom-up” gamification, we can exemplify how it can be developed. This provides insights into aspects that need to be considered in the design process. We decided to create a task management application (called *BU-ToDo*). This seems reasonable given the derived requirements (e.g., **R1** in Table 4.2) and from a domain-independent viewpoint, as aspects that might be gamified can be considered as a task (see **R12**). Furthermore, such an application does not require particular knowledge and can in theory be used by everyone.
- Goal_{BUToDo} 2** *Qualitative evaluation of “bottom-up” gamification:* *BU-ToDo* is used in a user study to not only validate the “bottom-up” concept but also to investigate which effects it has on the users themselves. Now, as an actual realization is used, it allows us to further investigate the perception of “bottom-up” gamification (e.g., to learn whether participants still assess the idea as reasonable when they are directly confronted with it).

In the following section, *BU-ToDo* is presented. As we based it on the found requirements (see Table 4.2), we reference these were applicable.

4.3.1 Concept and System Design of *BU-ToDo*

In contrast to existing gamified task management apps which were investigated scientifically, such as *Task Hammer* and *Epic Win* [136] or *HabitRPG* [138], the interface and experience of our system *BU-ToDo* is not (necessarily) game-like, nor do we have a fixed theme (such as a role-playing context as in *Epic Win*). Both were done to account for users who might decide they do not want to use gamification, following the “bottom-up” definition. Nonetheless, we offer the option to make the app more game-like visually, as the users’ scores or achievements can be visualized on the dashboard.

Users can, for every task, decide whether they want to use game elements or not. If users want to use such, they can decide – for every task – which they want to use (**R8**): several elements are offered and users can combine them as they see fit (see below). Besides the above mentioned related apps, there are other apps that gamify task management (e.g., *LevelUpLife*, *Stikk* or *ChoreWars*). To our knowledge, these have not yet been investigated scientifically and they vary in their degree of customization options. Compared to our approach, they are offering less flexibility concerning the game elements (i.e., amount of game elements, customization options per element and combination options), i.e., these are not offering “bottom-up” gamification following our definition.

Task Management Features

An informal review of 30 task management apps shows a large variety of functions. To our knowledge, no analysis reports which features are beneficial. As solving this was not our focus, we implemented a core set of functions that seems reasonable for managing tasks. The app was created as a mobile web app, to ensure better compatibility across different devices. An overview of tasks is used as the primary view (see Figure 4.2, left), which is sortable and filterable. Here, every task is depicted with its name and category, its due date and whether it is reoccurring. New tasks can easily be created by clicking on the plus sign. A new task (see Figure 4.2, right) can be named, described, categorized, prioritized and a due date can be added. Additionally, tasks can be set to be reoccurring, and reminders can be added and configured. To allow an easy and fast creation of a task (**R2**), only the name is a mandatory. Optionally, during task creation, users also have the chance to add game elements to the tasks as they see fit (**R8**). Tasks can be edited and can be checked off after they have been done. Reminders are provided as e-mail notifications (**R5**). Concerning **R10**, we decided not to allow customization of notifications, to ensure that all were received in a later study setting. In general, only social element requests (see below) and reminders (that the users can set themselves) produced notifications (via e-mails), i.e., the amount of potential notifications can be considered as limited. Details to game elements are hidden behind a “more” button to avoid cluttering the interface, especially when several are selected for a task.

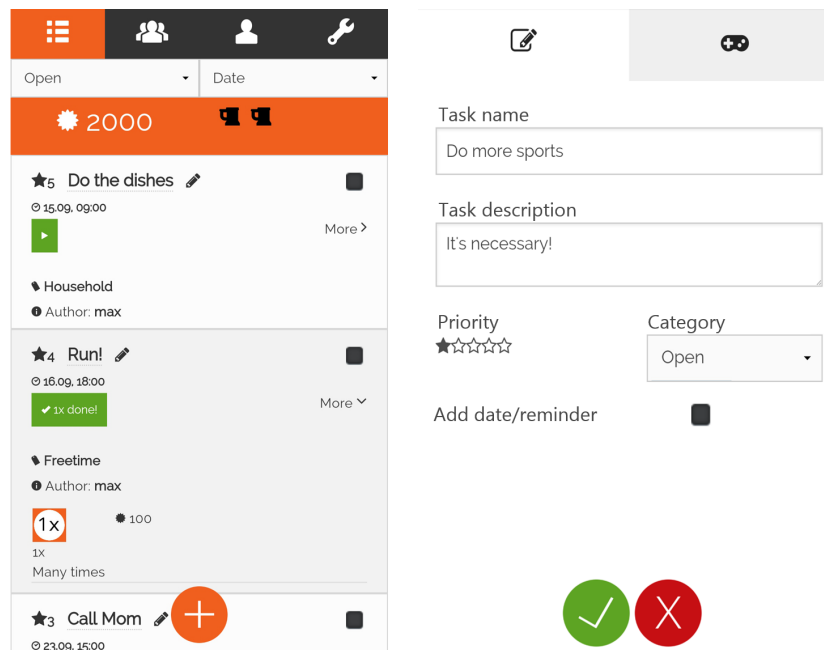


Figure 4.2: *BU-ToDo* main task user interface. Left: Task overview screen. Right: Task creation screen.

Integrated “Bottom-Up” Elements

Users can decide, on a per-task level, which game elements they want to use for a specific task. Users can thereby always decide how much pressure they want to put on themselves (R9). In the element selection these are grouped as “goals”, “rewards” and “play together” elements (see Figure 4.3, left). If the user wants to use game elements for a task, we only require that the user defines at least one goal and at least one reward (mainly a programmatic restriction). Every attached element can be further configured depending on the corresponding element (see Figure 4.3, right). While points, for example, only allow a user to set their amount, other elements allow more sophisticated setup options. Additionally, some game elements, when selected for a task, add further interface elements to the task overview (placed next to this particular task). For example, the timing-related game elements would add a timer that can be started or stopped by the user (see play button in Figure 4.2, left). Overall, we do not limit the number of elements that can be added to a task. Thus, users are also able to add several game element combinations. For example, a user could set up the task “*Do the dishes 5 times this week, with one set needing to be done in under 5 minutes. If completed, receive 200 points and a badge*”. The task would appear in the overview and the app would offer buttons to measure the time, helping the user to keep track. In this example, *what should be done, how often, how fast, the amount of points and the appearance and name of the badge* could be defined by the user. Thus, users can define the “content” of the elements, their composition and the elements themselves.

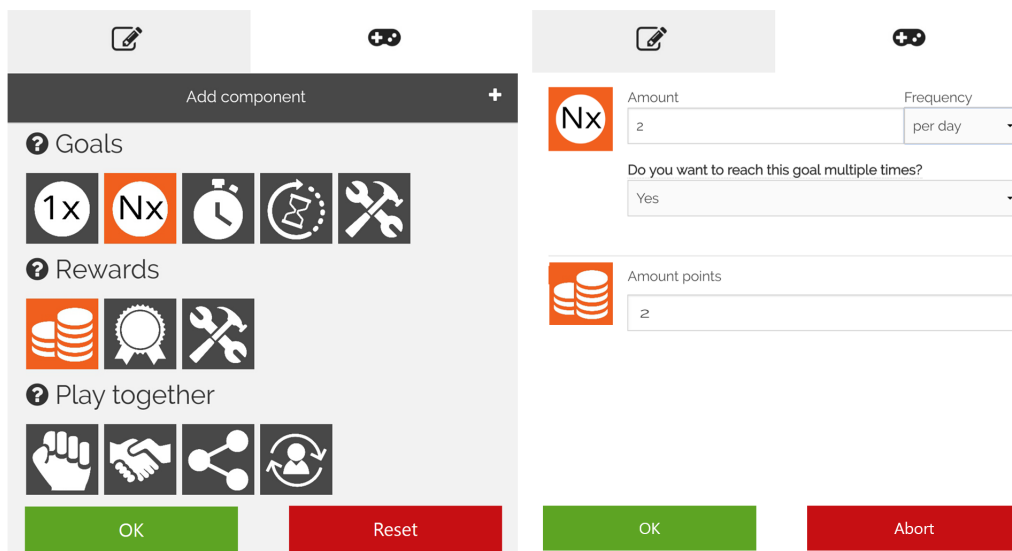


Figure 4.3: Game element configuration screens in BU-ToDo. Left: Game element selection screen. Right: Customization on a per-element level.

We followed the game element review (see Section 4.2.1) to decide which elements should be available. To allow social features, other users can be befriended (**R4**) and as long as it is permitted by this user (**R11**), others can inspect their friends' profiles to see their points and received achievements. We did not include the *social recognition* feature (i.e., informing when a specific task was started or ended), as this was not perceived well in our online study. We did not integrate *story elements*, nor the *virtual goods* or the *avatar*, as we judge those as too much effort for users (especially considering **R2** and the fact that participants disagreed to the statement that they would be willing to add new content into the system). We also did not integrate *assistance tools* (e.g., a graphical tool to create virtual goods) into the prototype, as they would shift the focus away from the core task management app (**R13**). To avoid introducing an artificial bias, we also omitted **R15**. Overall, the following elements were available (see also Figure 4.3):

- **Goals:** Users are able to specify that tasks need to be done once or multiple times. Selected rewards are only then unlocked. A time frame can be specified, and users can set up goals to be reachable several times (e.g., receiving a reward whenever the goal was fulfilled twice a week). Users can indicate that they have done a task once by clicking on an additional button in the main screen. They are also able to specify that a task needs to be done for at most or at least a certain duration (**R6**). As mentioned, a timer in the app is then shown and can be used for this. Finally, a user can also define a custom goal (e.g., “Lose 10 pounds”) which users need to manually mark as achieved. This category fits the *clear goals* gamification element. We also added progress bars (*progression*) and provided *feedback* for tasks that need to be done multiple times (**R3**).

- **Rewards:** Users can assign *points* for goal fulfillment. The amount is up to the user (and can be negative). Another reward we offer are *achievements* (R7), which can have an icon (i.e., badges) and/or a textual representation (e.g., “Master Runner”). Icons can be selected and colored as users see fit. *Self-defined rewards* can also be defined (e.g., “Buying ice cream”).
- **Play together:** The app offers social features (R4): a user can invite friends to tasks and start a *competition*. Their performances are visualized on *leaderboards* with metrics that can be adjusted by the task creator (e.g., points or time) to make rewards comparable here. Friends can also *cooperatively* handle tasks and can accumulate points together. How much users contributed is then shown for each task. Additionally, for every task, it is possible to assign a *reviewer* that checks that the task has been done properly, and only then a reward is unlocked (R14). Friends can assign tasks to each other with rewards that they have defined. In general, for every invitation type (i.e., either a task assignment or an invitation to a competition), invited users always have the option not to participate, to increase the perceived freedom of choice (R9).

4.3.2 User Study with BU-ToDo

We conducted a user study to analyze how people use *BU-ToDo*. This study was mainly exploratory, which is why we did not formulate hypotheses. The goal of this study was to find out how the “bottom-up” elements are perceived and whether people appreciate the autonomy offered in such a system, as well as to potentially learn further aspects to be considered in such approaches.

Method

German participants were recruited via word-of-mouth communication. The only task we provided them was that the application should be integrated in their daily life as they saw fit, but they should use the app at least once per day. People who agreed to participate were provided with a link to a pre-session questionnaire that assesses gaming affinity and how they currently manage their tasks. After completing it, they were asked to watch an online tutorial video, explaining the app and its game elements. The video contained the link to the app, i.e., they only received access after having watched the video. After six days the participants received a link to a mid-session questionnaire, assessing their perception of the app. After twelve days, a post-session questionnaire was provided, assessing their experience with the app and the gamification elements offered. All questionnaires consisted of a mix of 5-point-scale questions with labels on every option (*disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, *agree*), yes/no questions and (optional) free-text questions. Additionally, we logged all interactions with the app to receive quantitative data.

Results

20 participants (ten male, ten female; age: <21: 3; 21–30: 15, 31–40: 2) participated. Two were school students, two apprentices, nine university students and seven employees. The sample classify themselves as having high gaming affinity ($M=3.9$, $SD=.8$, $Mdn=4$) and they (partially) could imagine solving private/work tasks playfully ($M=3.5$, $SD=1$, $Mdn=4$ / $M=3.2$, $SD=1.2$, $Mdn=3.5$). 17 (85%) reported using analog or digital tools for task organization, with nine (45%) using handwritten to do-lists and six (30%) using apps, but they all agreed that a task management app could help them to better organize their tasks ($M=4.1$, $SD=.7$, $Mdn=4$). 10 participants (50%) a priori thought that game elements could motivate them to have more fun solving tasks, while 14 (70%) thought that tasks would be solved more efficiently with them.

We removed tasks that were obviously meaningless (e.g., “TestTestTest”), and considered only participants who completed all questionnaires and who did not deviate by more than one standard deviation downwards in two of these measures [197]: number of tasks created, task interactions³⁴ or general interactions³⁵. This led to the exclusion of two participants. From their answers, it remains unclear why they had so few interactions. For the social features, we ensured that participants knew at least one other participant (and added them as friends in the app). After data cleaning, this goal was not reached, as the exclusion led to one user without friends in the app. On average, users had 3.5 friends in our app, with seven being the largest friend count, achieved by three users.

General app usage: 199 tasks were created, with 11 tasks per participant on average ($SD=5$). We counted on average 11 ($SD=5.3$) general and 4 ($SD=1.1$) task interactions with the app per day. Even though these numbers seem low, it needs to be kept in mind that participants were requested to use the app with real tasks they would normally add to such a list. A similar task creation rate was reported in [15]. On average, 16.6 tasks ($SD=12.9$) were created by all participants every day. Comparing the activity level in the first with the last six days, fewer tasks were created (157 vs. 42). However, the average amount of task interactions remained stable (4.1 vs. 3.9), indicating that even though lower amounts of new tasks were added, the interactions with the app remained similar. Two reviewers inspected the tasks separately and categorized them. The classification revealed that significantly more private (71%) than work tasks (19%) were entered, as a paired t-test showed ($t(17)=5.9$, $p<.01$). The remaining 10% could not be classified specifically. The app was used not only for common tasks (e.g., “Gas up today”), but also for behavior change tasks (e.g., “Eating salad three times a week”). A precise quantification of the latter is not possible; from a title alone it is often not clear whether a behavior change or a one-time task is meant. Nonetheless, in general this showed that participants also added behavior change tasks in the application as asked for in our previous online study (see Section 4.2.2).

³⁴ That subsumes task execution, finishing a task, starting/stopping a timer.

³⁵ All potential interactions with the app. It subsumes task interactions.

13 of the participants (72%) stated that they generally want to use the app, after it has left its prototype status. A third of the sample could already imagine using the app as it was subsequent to this study. These are promising results towards the app design. From a usability perspective, the average *System Usability Scale* score [29] of the app was 72.4 (SD=11.6, Mdn=72.5), which is considered as acceptable [10]. We also explicitly asked participants whether the adding of tasks or the addition of game elements was cumbersome. The sample overall disagreed to both (M=1.6, SD=.9, Mdn=1/M=2.7, SD=1.2, Mdn=2). Participants stated what they liked in the app: six times the graphical appearance and conceptual aspects were highlighted positively, and four participants explicitly stated that the app had good usability.

“Bottom-up” gamification usage: Participants created significantly more gamified tasks³⁶ than non-gamified ones, as a paired t-test revealed (142 vs. 57; $t(17)=4.3$, $p<.001$). We found no significant difference in the number of gamified work (76%) and gamified private (68%) tasks ($p=.39$). The ratio of gamified to non-gamified tasks created in the first and last six days also did not differ significantly ($p=.36$), i.e., when a task was entered, it was more likely gamified.

Gamified tasks were equipped with 2.9 game elements on average (SD=2.5), with 70% of the gamified tasks having exactly two game elements (the minimum our app requested), indicating that participants tended to stick with simple elements instead of using complex combinations. Table 4.4 shows that there were two favorites: *solving the task once/multiple times* seems easily explainable because of the nature of tasks in general. The second most often used element was the reward *points*. Compared to the other reward types (the app requested a reward to be enabled for every gamified task) points were easy to use, as only an amount needed to be entered. In contrast, badges needed more effort, and the custom reward would not alter the app after unlocking. Participants were also asked to rate how motivating elements were, and could provide free-text remarks (both also visible in Table 4.4). The table shows that element use and perception do not necessarily correspond, as for example competition is assessed as motivating by a majority of participants, but was only used twice. A reason for this is that for many elements participants stated that they had not had a suitable task during the study where they could use this element. When counting game elements that were used more than once (29 times an element was only used once), a participant used on average 3.8 different game elements (SD=1.9, Min=0, Max=8), with only one participant having used no game elements at all. In general, this indicates that participants tended to use the same core elements in the app (corresponding to the online study; see Section 4.2.2), but would also use other elements, depending on the scenario. To what degree cannot be derived from the data of this study. The sample also subjectively agreed to the statement that they used the same elements throughout the study (M=4.1, SD=.9, Mdn=4). Even though the social features were not used often, the sample agreed to the usefulness of these in the context of the app (M=4.2, SD=1.3, Mdn=5).

³⁶Tasks with at least one game element.

Game element and motivational assessment	Times used	Main reasons for not having used the game element or not finding it motivating
Goal: Solving the task once or multiple times (44%)	119 15	None of my tasks would have fit for the multiple time element (6×)
Goal: Doing the task within or for some time (45%)	33 24	None of my tasks would have fit (7×)
Goal: Setting a custom goal (16%)	3	None of my tasks would have fit (5×); formulation as task, not a goal in the task (2×); other elements were sufficient (2×); why should this motivate me? (2×)
Reward: Points (61%)	108	No further usage options (3×); no comparability option to others (3×)
Reward: Badges (66%)	48	No comparability option to others (1×)
Reward: Custom reward (16%)	3	None of my tasks would have fit (5×); no effect in app, thus no need to enter it (4×)
Social: Sharing the task (50%)	21	None of my tasks would have fit (1×); not seen (1×)
Social: Set a reviewer (56%)	9	None of my tasks would have fit (1×); receiving a task is not motivating for me (1×)
Social: Collaboration (39%)	5	None of my tasks would have fit (5×); compared to task, too much effort (2×)
Social: Competition (61%)	2	None of my tasks would have fit (2×); too complex (2×); compared to task, too much effort (2×)

Table 4.4: Game element usage in the study: number of uses, percentage of participants (in parentheses) who perceived the element as “somewhat motivating” or “motivating”, and answers on why the element was not used.

Participants agreed that choosing game elements for themselves was appreciated ($M=4.5$, $SD=.8$, $Mdn=5$). For selecting a reward for themselves, the sample still tended to assess this slightly positively ($M=3.5$, $SD=1.4$, $Mdn=4$). A surprising result was that the sample was indifferent on whether they could have forgone the game elements ($M=3.1$, $SD=1.2$, $Mdn=3$). One explanation for this might be the prototype state of the app, but it can also be an indication that even though selecting game elements is perceived positively, not everyone likes the additional effort involved. Concerning the subjective effects “bottom-up” gamification had on participants, three reported that they thought they solved tasks more efficiently with the game elements, seven thought that they had more fun solving tasks, seven thought that they were motivated to solve the tasks sooner and five reported solving tasks more consciously. Overall, twelve participants (66%) agreed to at least one of these statements, showing that the presence of game elements subjectively had an effect. Further support for **R15** is established, as our sample wanted to see how other users gamified their tasks ($M=3.9$, $SD=.9$, $Mdn=4$) and wanted to directly copy these elements ($M=4.1$, $SD=.9$, $Mdn=4$). In general, this indicates that such a system might also provide the chance to enable participants to easily inspect and utilize complete game configurations, instead of just providing the core mechanics to create configurations.

Half of the participants mentioned the gamification aspect as a positive feature of the app (in particular the easy self-creation of badges, combination options for game elements, offered reward options or individual highlighting of game elements) and nine times the social aspects (solving tasks together, assigning tasks to others, option to let tasks be reviewed). The most often mentioned negative aspect was the lack of an option to edit game elements (11×), which was not implemented, as we assessed it as also easy to re-create a task with the updated elements. From the comments, we learned that this is not only a usability suggestion. Participants would have used this to adapt rewards for tasks that turned out to be easier or harder. This further adds to the “bottom-up” idea. Points were another important topic that was mentioned negatively (7×): reviewing points in other users’ profiles was not comparable, due to the “bottom-up” approach. Suggestions to overcome this focus on functionality in which the app automatically derives points. One user proposed the following suggestion: based on user-assignable difficulty levels, points should be derived, and could also be individualized by asking a priori how hard specific task categories are for the user, e.g., doing sports could be easy for a certain user and would lead to fewer points. Another aspect reported that related to points was that they should be usable for something, e.g., buying predefined, virtual goods or everyday rewards. Both of these suggest that there might be parts in which a “bottom-up” approach might be interrupted for the sake of reasonable usage of some elements, or that a mixture of “bottom-up” and “top-down” aspects can be beneficial.

Discussion

The study provided different pieces of evidence that the increased autonomy offered in the “bottom-up” task management application is appreciated by the participants. This is in line with the online study (see Section 4.2.2), in which “bottom-up” gamification also was perceived positively. This was our initial hypothesis, given the related work on autonomy (see Section 2.2) and customization (see Section 2.3.2) in gamification. The various game elements offered were seen subjectively as motivational to a certain extent by the participants. Interestingly, variance across the game elements used in the tasks was lacking (which was also similar to the online study) and participants tended towards simple gamification elements, instead of using more complex combinations. This can be accounted for by the short runtime of the study, as many participants stated they did not have a suitable situation for using the element. Another explanation for this is that in “bottom-up” scenarios complex game configurations are not necessary (although we will see in Section 4.5 that users can develop more complex ones) and/or that the participants simply were satisfied with their initial ideas. As this study was meant to assess the perception of “bottom-up” gamification, we focused on an application which solely consisted of “bottom-up” elements. The suggestions for how to handle points, the criticism that points are not comparable (even though they would be in a competition) as well as the nearly consistent answer

that participants wanted to see how other people used game elements for tasks, and wanted to be able to utilize the same ideas, showed that certain “top-down” elements in “bottom-up” gamification might be relevant. There is also the chance that some users actually want to make use of gamification without bothering with customizations (even though our studies suggest otherwise). Potentially, these approaches should not be seen as mutually exclusive. Importantly, two-thirds of the participants reported that they had more fun completing the task, or finished it earlier, more consciously, or more efficiently through the game elements (or a mix of these aspects). Based on the nature of the app, those aspects were not easily verifiable from an objective point of view, as we would have needed to be able to monitor all potential tasks that a user might put in the app (e.g., to measure how efficient he or she was). Thus, in this study, we can only state that “bottom-up” gamification appears to motivate users subjectively, and it indicates that it fulfills the gamification goals to make activities more fun and engaging [64] (at least in the short term). Because of the short runtime of the study, other explanations could also be framing [176] (e.g., the knowledge to having a game can impact the motivation positively) or novelty effects [107, 148, 332], also known for “top-down” gamification. If those are the explanatory factors, it would at least show that “bottom-up” gamification elicits similar effects as “top-down” gamification, but at the same time offers more autonomy to the users.

This study had limitations, primarily the short duration and the low number of participants. Keeping in mind that this study was the first prototype-based exploration of “bottom-up” gamification, we see this as acceptable. A further limitation can be seen in that we only used task management context for the approach. Although an informed decision, it is questionable whether “bottom-up” gamification is perceived differently in another “host app”.

4.3.3 Contribution to the Thesis’ Questions

Overall, the online study (see Section 4.2.2) and the above presented study with *BU-ToDo* added to **RQ2** (see Section 1.4). With the online study, we assessed participants’ views towards “bottom-up” gamification without a specific prototype (**Goal_{BUExpectations}**), which were replicated in an “in the wild” setting. Taken together, this provides a first holistic view on “bottom-up” gamification, showing that it is reasonable to follow its idea introduced in Section 4.2. Considering **Goal_{BUToDo} 1**, we showed how “bottom-up” gamification can be established in an application that considers arbitrary tasks. It is easy to adapt this setting to specific applications, as we hypothesized that other applications also have one or more specific tasks that the user tries to fulfill. Thus, our application can be considered a general example for how to apply “bottom-up” gamification. Considering **Goal_{BUToDo} 2**, through a user study, we learned that participants perceive the “bottom-up” approach as valuable (even though more thought and effort needs to be invested by users), potentially, as it offers more autonomy. In addition, this study revealed that the full range of customization might not be

necessary and that “top-down” elements might be interesting in a “bottom-up” setting as well: either to reduce the effort the user needs to expend, or to make some elements more relevant in the setting itself. Participants in the user study also reported positive subjective effects on motivation. Finally, although we were able to show that individuals seem to use the same elements across scenarios, we also saw that all elements were used in the application (although not to the degree visible in Section 4.2.2). Concerning the thesis focus, the application and the study showed that allowing users to influence the motivational aspects a system offers seems worthwhile. As explained above, based on the nature of the application, we were not able to measure the positive effects reported within this study with an objective measure. In the next section, we will present a system that utilizes the “bottom-up” setup presented here, together with the microtask idea of *ExpenseControl* (see Section 3.2). This allows us to measure the effects quantitatively (i.e., by measuring how many microtasks were solved with “bottom-up” gamification in comparison to a baseline), but will also allow us to make comparisons to “top-down” gamification. The latter is specifically interesting, as work such as [106] indicates that there are (“top-down”) gamification approaches that only provide mixed results, posing the question of whether users also might decide wrongly for themselves. If “bottom-up” gamification provides positive objectively measurable results that are “better” than “top-down” approaches, this would indicate that “bottom-up” gamification might be a valuable alternative.

4.4 “Bottom-Up” Gamification in a Microtask Setting

Although the task management application was useful to explore the general view of “bottom-up”, it had the issue that it appears difficult to evaluate the effects “bottom-up” gamification elicits quantitatively. Thus, so far, it is not clear whether self-tailoring also has positive effects from a quantitative viewpoint, similar to the effects that were reported for “top-down” gamification (e.g., [272]). Thus, we had the following goals:

- Goal_{BU_{Micro}} 1** *Creation of a “bottom-up” gamification application offering quantitatively measurable outcomes:* To be able to investigate the objective effects “bottom-up” gamification elicits, we need to develop a system that has a quantitatively measurable dependent variable. This system should also offer different self-tailoring gamification degrees, as comparisons seem valuable given the results gained in Section 4.3.
- Goal_{BU_{Micro}} 2** *Quantitative evaluation of “bottom-up” gamification:* Through the application resulting from the previous goal, we are able to investigate the effects “bottom-up” has quantitatively. In addition, we can investigate different degrees of “bottom-up” gamification, in comparison to a “top-down” and a no-gamification approach.

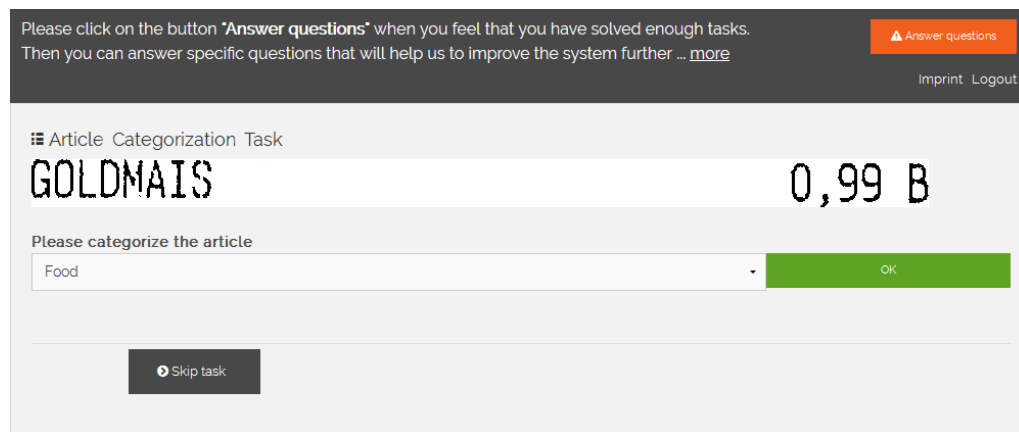


Figure 4.4: “No gamification” interface of the *BU-Microtasks Platform*.

4.4.1 Concept and System Design of the *BU-Microtasks Platform*

To account for **Goal_{BU-Micro} 1**, we decided to combine the “bottom-up” gamification part of *BU-ToDo* (see Section 4.3) with the microtasks of *ExpenseControl* (see Section 3.2.1). The amount of microtasks solved in such a system would be the dependent variable in a study setting. We have already seen in Section 3.2 that gamification in general encouraged people to do more tasks. To be able to reach a broader user base and to minimize potential bias, we created an online platform in which just the tasks are presented, i.e., we did not integrate the “bottom-up” concept in the digital accounting book context. As users do not have further benefits, we hypothesized that this affects the overall number of solved microtasks, but not the relative effects between using gamification and no gamification.

The prominent part of the platform is the microtask area. Here, we present the microtasks similarly as in *ExpenseControl*, and also used the three microtask types (*Classification*, *Article correction*, *Article categorization*) (see Section 3.2.1). The platform uses the more than 6000 different pictures that were generated during the user study of *ExpenseControl* (see Section 3.2.2). Participants can skip tasks instead of solving them. Above the microtasks area, a button led to questions to be used in a later study. The platform was created in German.

To be able to compare gamification approaches, the platform implements five different conditions. With “no gamification” being a baseline (see Figure 4.4), in which only the microtasks are presented, the other four conditions model different gamification approaches. Besides the “bottom-up” condition (which is based on the approach taken in *BU-ToDo*) and a “top-down” condition (which is based on the gamification used in *ExpenseControl*) we also added two conditions that are in between this spectrum. This accounts for the finding that “pure bottom-up” gamification might not be optimal for some users (see Section 4.3.2): in “selective top-down” users have the option to select which game configuration they want (without giving them the option to edit or add a new configuration) while in

The screenshot displays the BU-Microtasks Platform interface. At the top, a dark header contains icons for a menu, user profile, friends, and settings. Below this, an orange bar shows a sun icon and the number '400'. A dark grey bar below the orange bar contains the text: 'Please click on the button "Answer questions" when you feel that you have solved enough tasks. Then you can answer specific questions that will help us to improve the system further ... [more](#)'. To the right of this text is a button labeled 'Answer questions'. Further right are links for 'Imprint' and 'Logout'.

The main content area is divided into three sections:

- Article Correction Task:** Shows the article 'FLEISCHKAESE' with a score of '1,23 B'. Below this is a text input field with the prompt 'Please name the article (without its cost)' and an 'OK' button. A 'Task counter: 40 (reset)' is displayed, along with a 'Skip task' button.
- Game Elements:** A list of game elements on the right side:
 - 1x Competition (10 points)
 - 1x Crowd Specialist 1 (1x icon)
 - Nx Crowd Specialist 2 (60x multiplier)
 - Nx Crowd Specialist 3 (120x multiplier)
 - Nx (180x multiplier)
 A 'Show all elements' button is at the bottom of this list.
- Highscores:** A table showing the top two users:

#	User	Points
1	ZFP2	3060
2	kassen	2830

Figure 4.5: “Top-down” gamification interface of the BU-Microtasks Platform.

“selective bottom-up” the game configurations can be selected and edited further (but without the option to add new elements to the configurations). Thus, these approaches modeled different degrees of customization. The conditions and their changes to the platform’s user interface are:

“Top-down” gamification: Users in this condition are confronted with a “top-down” defined game setting, similar to *ExpenseControl*: we award 10 points for every solved microtask, a leaderboard shows the performance in comparison to others, and we integrated the *Crowd-Specialist* and *Task-Specialist* badges. The leaderboard is initialized with 10 entries (ranging from 0 to 1000 points). To ensure comparability, only other users in the “top-down” condition are shown in this leaderboard. The points are visible at the top and the leaderboard is shown below the microtask area (see Figure 4.5). The game elements are placed on the right side and the available badges can be inspected further. Additionally (similar to all other gamification conditions), users are shown a header, in which they can inspect their recent badges and points, befriend others and view their profiles. The users in this condition cannot disable, edit or change any of the game elements.

No.	Name in drop-down	Game setting description
01	Points	10 points per solved task.
02	Points, Leaderboard	10 points per solved task, and performance is shown on a leaderboard.
03	Points, Badges	10 points per solved task and access to badges as explained in the “top-down” condition.
04	Points, Time pressure	10 points per solved task, if the solving time was below 5 seconds. A timer is shown in the microtask area and starts at 0 whenever a new task is shown.
05	Points, Leaderboard, Time pressure	10 points per solved task if the solving time was below 5 seconds. A timer is shown in the microtask area. Performance is shown on a leaderboard.
06	Points, Leaderboard, Badges	10 points per solved task, performance is shown on a leaderboard and there is access to badges as explained in the “top-down” condition.
07	Points, Cooperation	1 point per solved task, and a list showing how many tasks were solved by the different users in this configuration. This is also visible below the microtask area.
08	Points, 20 Tasks, Leaderboard	200 points are assigned for every 20 solved tasks, and performance is shown on a leaderboard.
09	Points every 2 minutes	200 points are assigned for every two minutes the user stays in this condition.
10	Points, 20 Tasks, Time pressure	200 points are assigned when 20 tasks are solved in under two minutes. A button is shown, with which a timer can be started. The timer runs down starting at two minutes.

Table 4.5: Available game configurations in the selective conditions of the BU-Microtasks Platform.

“Selective top-down” gamification: Users can select which “top-down” defined configuration they want to use and can switch it anytime. The user interface is similar to Figure 4.5. The only difference is a drop-down menu below *Game Elements* in the right part of the interface. It contains ten configurations (see Table 4.5) which are based on commonly used elements in crowdsourcing [206] and in general [106, 134, 269]. After a selection, we show an explanation of the configuration and the corresponding game elements, similarly to the “top-down” condition. Additionally, a help button opens a dialog with an explanation of the game element icons. The initial configuration is selected randomly for every user. Users cannot decide not to use any game configuration, leading to an always active gamification. The game elements are not editable and no new game elements can be added to any configuration in this condition. All configurations that contained leaderboards are treated separately, i.e., points gained in one configuration are not visible on leaderboards in other configurations, as the necessary action to receive the points varies and is thus not comparable. The same is true for badges, which is why we need to assign them “per configuration” anew (unlocked badges remain unlocked if a user decides to come back to this configuration later). All users in this condition are directly shown with their user name on these leaderboards. All other parts are as in the “top-down” condition.

“Selective bottom-up” gamification: Users have the same interface and options as in the “selective top-down” condition. The difference here is that all configurations can be edited further. For configuration 05 (see Table 4.5), a user could adjust the points from 10 to, for example, 123 and the time from 5 seconds to 11 seconds. The only limitation here is that users are not able to add further elements to configurations or to create new ones. A small edit icon (being the only difference in the user interface compared to the “selective top-down” condition) is always shown next to the active configuration and leads to a dialog in which these adjustments can be made (similar as in the task management application). As a consequence of this freedom, leaderboards/cooperations are treated differently to “selective top-down”, as points are not necessarily comparable anymore. After changing to a configuration with one of these elements, users can invite friends, and only these are shown on the leaderboard/cooperation board. Only the “host” can further customize such configurations, i.e., it is not possible for friends to alter a configuration where they are taking part by invitation, but they can simply create a new one anytime and invite others themselves. An important difference in this condition is that users started without an active game configuration and can switch back to this state whenever they want. Thus, users have the freedom to not use any gamification in this condition.

“Bottom-up” gamification: This condition offers the most freedom for its users, as game elements can be combined freely. We used the elements of *BU-ToDo*, but removed the custom goal and reward as well as the reviewer element, and added a time element (see Figure 4.6) which allowed users to receive a reward periodically (e.g., every five minutes they are on the platform). We allow multiple configurations to be active in parallel and all elements can be edited at any time. For comparability, every configuration that is offered in any of the other gamified conditions can be created here as well. As explained in the “selective bottom-up” condition, the leaderboards/cooperations also need to be friends-specific here. Thus, for every configuration in which such an element is used, a separate leaderboard/cooperation is created to allow comparisons. Friends can be invited to participate in this configuration (with the same restrictions as already explained above). As several configurations with leaderboards and cooperations can be active in parallel, we also allow users to customize which of these are directly visible below the task area. Users remain in such leaderboards (whether visible or not) as long as they want, i.e., they can solve tasks and also improve their score for currently-not-shown leaderboards as well. Initially no game elements are configured; thus, users again have the freedom to not use any gamification at all in this condition.

4.4.2 User Study with the *BU-Microtasks Platform*

We conducted a study to target **Goal_{BU-Micro} 2** by using the presented platform. We had the following hypotheses:

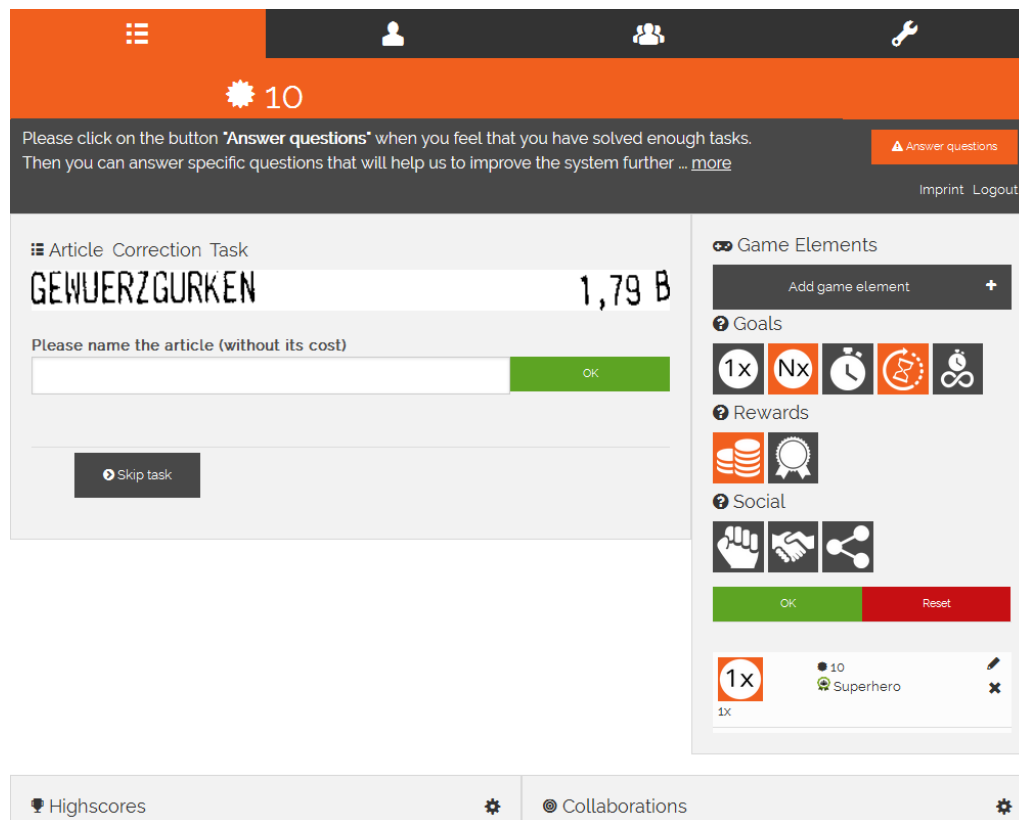


Figure 4.6: “Bottom-up” gamification interface of the *BU-Microtasks Platform*.

- H1** Participants in the gamified conditions will solve more tasks than participants in the “no gamification” condition.
- H2** Participants in the conditions offering customizable gamification (“selective top-down”, “selective bottom-up”, “bottom-up”), will solve more tasks than in the “top-down” condition.

H1 is supported by work showing that gamification is beneficial (see Section 1.2.1), which was also seen in *ExpenseControl* (see Section 3.2.2). **H2** is based on our previous studies’ results (see Section 4.3.2) and the work reported in Sections 2.2 and 2.3.2, i.e., that offering more autonomy and choices is beneficial for motivation.

Method

On the platform’s start page we briefed participants with the cover story that they will solve microtasks which will improve the recognition performance of an algorithm for automatic digitization of receipts and that they can help by participating. We also put emphasis on our research direction, i.e., that we want to create recognition algorithms that could later be used in easy-to-use digital

household accounting books. This should lead to a general positive framing effect, as shown in [198], which we deemed necessary considering the finding that the microtasks were not engaging (see Section 3.2.2). We did not mention that gamification elements are integrated, to avoid attracting only gaming-affine people or giving the impression that we have a game scenario, which could also affect the perception of the system [176].

When participants decided to take part in the study, they needed to enter a user name and a password (but no further data was requested). We clarified that with these credentials they could leave the platform and continue anytime later. This was followed by demographic questions, a question on their affinity to games, and a small tutorial in which every task type was explained and three such tasks needed to be solved. Besides the rationale of giving an introduction to the task, we also used this to learn how these are perceived: after each triple, we let participants answer the statements “*Solving these tasks was fun*” and “*It was easy to solve these tasks*”. Additionally, we asked how many such tasks they could imagine solving at a time and in a week. We provided drop-down menus in which they could simply select a range (for the first one, in 10 segments, ranging from zero to “>40”; for the second, 10 segments ranging from zero to “>80”). All questions that required a response to a statement had to be answered on a 6-point scale to force a decision, labeled *strongly disagree* to *strongly agree*. Up to this point, every participant had seen the same content and was now assigned to one of the five conditions. After the tutorial, participants were allowed to enter the user names of friends, provided they were not in the “no gamification” condition. We stated that certain elements are integrated that might make the solving of tasks more interesting and that some of them could be done with friends. When a new participant entered user names of friends, we ensured that he or she would be assigned to the condition that most of his or her friends were in. If no user names were provided, we distributed new participants to the conditions equally.

Subsequently, participants received access to the main view. A guided tour explained all areas of the interface. The explanations for the gamification aspects were adapted for the different conditions. Tasks were presented in the same order to ensure that no ordering effects might be a reason why some participants discontinued earlier than others. While this is reasonable for our setup, it is not advisable in a general crowdsourcing setting as recent work shows [32, 217]. Additionally, we strictly alternated between the three task types to introduce variance. The number of solved tasks, the task correctness rate and the task solving time represent our dependent variables (with the different conditions being the independent ones). Participants did not receive any feedback after they had solved the task and the study formulated no upper limit on the task participants could/should solve. As a consequence, a post-session questionnaire was provided, but was treated as optional. With it, we had the goal to better understand what drove participation. Participants could fill it out anytime, even though we stated textually that it should be done as soon as they had the feeling of having solved enough tasks. The questionnaire itself consisted of two parts.

The first part assessed the intrinsic motivation, by using a German, twelve item, 5-point scale version [328] of the *Intrinsic Motivation Inventory (IMI)* [59], and the second part contained statements (on the same 6-point scale as above) that focus on the perception of the game elements. Additionally, we logged all interactions with the platform. The link to the platform was distributed via student mailing lists (consisting of design, computer science and psychology students) and social networks. We accompanied the link with a short explanation that participation would improve an algorithm to digitize receipts. Also here, no hint towards a motivational study was given so as to not introduce a bias here either.

Results

129 participants finished the demographic questions and the tutorials, but only 106 participants also solved at least one microtask. Only these will be considered in the results (48 male, 50 female, eight no answer; age: <21: 13; 21–30: 68, 31–40: 13, >40: 12). 82 (77%) characterized themselves as gaming-affine on the 6-point scale (answering with 4 or more). Only two participants befriended each other and only four re-visited the page and continued with solving microtasks.

Game element usages: We analyzed the gamification usage across conditions:

“No gamification” and “top-down”: The “no gamification” condition ($n=23$, tasks solved: $M=47.8$, $SD=43.5$, $Mdn=40$) did not offer any game components and the “top-down” condition ($n=22$, $M=61.6$, $SD=76.8$, $Mdn=39.5$) provided no options to change the game elements. For the latter, we have no indication whether the game elements were noticed in general.

“Selective top-down”: Of the 20 participants in this condition, six switched the initial configurations at least once. Overall, 42 switches were done by them ($M=7$, $SD=5.7$, $Mdn=6$). We analyzed whether people that switched their initial configuration solved more tasks ($M=82.2$, $SD=75.5$, $Mdn=60$) than those that stuck with their initial configuration ($M=30.3$, $SD=40$, $Mdn=15.5$). A Mann-Whitney U test revealed that the amount of solved tasks is significantly different in these two groups ($U=72$, $z=2.5$, $p<.05$, $r=.55$). Not switching the configuration could suggest that participants were not interested in gamification, did not notice these elements or that the initial configuration was already motivating for them. The latter seems unlikely, as the average number of solved tasks was lower than with “no gamification” and “top-down” gamification.

“Selective bottom-up”: Of the 20 participants in this condition, eight switched to a game element configuration at least once. Overall, 48 configuration switches were done by them ($M=6$, $SD=4.5$, $Mdn=5.5$). As editing was possible in this condition, we also checked how often an edit was actually done, but only three edits happened. We compared the number of tasks solved between participants who activated at least one game configuration ($M=77.8$, $SD=106.9$, $Mdn=45$) and those who did not use any ($M=22.3$, $SD=19.7$, $Mdn=15.5$) and found a significant difference ($U=75.5$, $z=2.1$, $p<.05$, $r=.48$).

Condition	Tasks solved					
	Total	M	SD	Correct _{Clas.}	Correct _{Cor.}	Correct _{Cat.}
Top-down (n=22)	1355	61.6	76.8	393 (85.6%)	161 (35.6%)	424 (95.5%)
Bottom-up (n=21)	1158	55.1	69	318 (81.1%)	127 (32.8%)	358 (94.5%)
Selective top-down (n=20)	917	45.9	56.5	210 (69.9%)	113 (36.9%)	297 (97.1%)
Selective bottom-up (n=20)	889	44.5	72.2	245 (79.5%)	85 (29.6%)	282 (95.9%)
No gamification (n=23)	1100	47.8	43.5	311 (82.3%)	148 (40.9%)	342 (95%)
Fixed (n=36)	1779	49.4	66.1	479 (80.1%)	201 (33.8%)	564 (96%)
Adaptable (n=23)	1892	82.2	87	525 (80.6%)	210 (33.9%)	591 (95%)
Without (n=47)	1748	37.2	40	473 (79.8%)	223 (38.4%)	548 (95.5%)

Table 4.6: Number of solved tasks in the *BU-Microtasks Platform* conditions. Correctness percentages for the different task types relate to the amount of solved tasks in this type. “Fixed”, “adaptable” and “without” are revised conditions based on whether participants had a choice and used it. More details on the re-partitioning are provided in the text.

“*Bottom-up*”: Of the 21 participants in this condition, nine used at least one game element. We compared the number of tasks solved between participants in this condition who used at least one game element ($M=86.3$, $SD=85$, $Mdn=61$) with those who did not ($M=31.8$, $SD=44.9$, $Mdn=20$) and found that this differed significantly ($U=84$, $z=2.1$, $p<.05$, $r=.47$). Overall, four participants only set up their game elements once, four used two game configurations (i.e., set up game elements, then changed them later on) and one used three. Only the cooperative element and the minimum duration time element (i.e., spending a certain time on a task) were not used. Overall, ten different configurations were created by these nine participants, indicating that users are quite diverse in what they think motivates them, but once this is selected, users stick with it; otherwise, more configuration switches might have been seen. This is in line with the findings gained in the previous study of *BU-ToDo* (see Section 4.3.2). Overall, the social elements (cooperation, competition, sharing) were not used often as only two participants were friends in the system and could use these reasonably.

Task solving across conditions: Concerning the ratings for fun (classification $M=2.9$, correction $M=2.8$, categorization $M=3.1$), no task type is particularly rewarding for most of the participants. Eight participants reported in their free-text answers that the tasks themselves were boring and variety was lacking. Table 4.6 shows how many tasks were solved in every condition. We conducted a Kruskal-Wallis ANOVA, but were not able to find a significant difference according to the amount of solved tasks and condition ($p=.92$). In the previous section, we showed that many participants actually were in a gamification condition, but did not use their offered choices as they did not set up game elements or used configurations. As different explanations for this are possible, we clustered the participants into revised conditions (see Table 4.6 for the amount of tasks solved in these new groups as well), according to the following schema:

- **“Fixed gamification (fixed)”**: As we have no indication whether participants in the “top-down” condition actually noticed game elements or not, they remain in their own group. We additionally add the participants that were in “selective top-down” but did not switch their initial group (see above for potential reasons why they might not have switched) and thereby were also effectively in a “top-down” setting (n=36).
- **“Adaptable gamification (adaptable)”**: All participants of the “bottom-up” condition who set up game elements and solved tasks, participants in “selective top-down” who switched their configuration at least once and participants in “selective bottom-up” who used at least one configuration were clustered. This group represents “bottom-up” concepts, with different degrees of freedom, that were used by participants (n=23).
- **“Without gamification (without)”**: Participants in “bottom-up” who did not set up any game elements, participants in “selective bottom-up” who did not switch to any game configuration and participants that were already in the “no gamification” condition were clustered into this group (n=47). The participants solved tasks without any gamification being active.

A Kruskal-Wallis test showed that the amount of solved tasks differs significantly between these new groups, $H(3)=10.5$, $p<.05$. A pairwise comparison with Bonferroni-corrected p-values showed that the groups “without” and “adaptable” ($p<.05$, $r=.34$) and “fixed” and “adaptable” ($p<.05$, $r=.46$) differ significantly. This hints that in conditions in which a user has a choice of how to use gamification, and uses it, more tasks are solved. We also compared the ratio of correct to incorrect tasks across the conditions (see Table 4.6), but were not able to find a significant difference with a Kruskal-Wallis test ($p=.66$ and respectively $p=.65$), i.e., no condition produced significantly different errors in the task solving. Analyzing the time spent to solve single tasks in the different conditions with the same statistical test reveals no significant differences in the initial conditions. The revised conditions revealed significant differences for the overall timings (“fixed”: $M=10.2s$, $SD=13.4s$, $Mdn=7.1s$, “adaptable”: $M=7.8s$, $SD=17.4s$, $Mdn=5.1s$, “without”: $M=9.7s$, $SD=14.3s$, $Mdn=6.8s$) at the $p<.05$ level. A pairwise comparison with Bonferroni-corrected p-values showed that participants in the “adaptable” condition spent less time per task than in the “fixed” and the “without” condition. The adaptable gamification approaches seem to have motivated the participants to solve tasks faster, without a significant loss of correctness.

Causality: The significant differences found in the “adaptable” condition might also be explainable by having users that were in general more open to participate *and thus* used the gamification options *instead of* them being the driving factor for solving more tasks. We considered the answers to the question for every task type after the tutorial (*“How many tasks would you solve in a stretch/in a week”*) condition-wise (initial and revised), but no significant differences were found ($p=.18$ for “at a time” and $p=.42$ for “in a week”), i.e., no condition had more participants that wanted to solve more/less tasks “a priori”. We also asked how

Condition	Intrinsic Motivation Inventory											
	Interest			Competence			Choice			Pressure		
	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn
Top-down (n=12)	2.3	1	2.2	3.6	.7	3.3	2.6	1	2.7	1.5	.5	1.3
Bottom-up (n=15)	2.3	.9	2	3.8	.8	3.7	3.2	1	3.3	2.1	.8	2
Selective top-down (n=12)	2.7	1	2.5	3.6	.7	3.3	3.3	.9	3.2	2.3	.9	2.3
Selective bottom-up (n=13)	2.7	1.1	2.7	3.8	.8	3.7	3.4	1	3.7	1.6	.8	1.3
No gamification (n=19)	2.5	.9	2.7	3.7	.7	3.7	3.3	.9	3	1.9	.9	1.7
Fixed (n=18)	2.4	.9	2.2	3.5	.6	3.3	2.7	.9	2.7	1.8	.8	1.3
Adaptable (n=21)	2.6	1	2.3	3.9	.7	4	3.4	.9	3.3	2	.8	2
Without (n=32)	2.6	1	2.5	3.7	.8	3.7	3.3	1	3.3	1.9	.9	1.5

Table 4.7: Results of the IMI in the BU-Microtasks Platform conditions.

relevant the topic of digital recording of receipts is. From the 70 participants who answered it, no significant difference could be found in the initial ($p=.14$) or the revised conditions ($p=.26$), but a trend was visible: The lowest values (i.e., finding it less relevant) can be found in “bottom-up” ($M=2.9$, $SD=2$, $Mdn=2$) and “adaptable” ($M=3.3$, $SD=2$, $Mdn=3$); the highest in “top-down” ($M=4.5$, $SD=1.4$, $Mdn=5$) and “fixed” ($M=4.2$, $SD=1.6$, $Mdn=5$). This hints that participants in “bottom-up”/“adaptable” were not more motivated by the topic itself. Additionally, we considered the answers to the IMI, which 71 participants answered (see Table 4.7). We conducted Kruskal-Wallis ANOVAs with the initial and revised conditions and the respective values the IMI reports. None of these values differs significantly across the initial/revised conditions (interest: $p=.77/p=.80$, perceived competence: $p=.63/p=.13$, perceived choice: $p=.24/p=.05$, pressure: $p=.16/p=.35$). Not finding a significant difference for *perceived competence* is expected, as the tasks (and the task selections) do not change during the experiment and also could not be adjusted with the game elements. The non-significant difference in *perceived choice* is surprising. Considering the average values reported in Table 4.7 (as $p<.05$ is almost reached), we see that in the revised conditions “fixed” seems most restrictive. The result for *pressure* could be explained by the absence of clear goals such as “You need to solve at least X microtasks”; thus, no source of pressure was available. As no group excels in the values for *interest*, the tasks seem to be perceived as equally uninteresting.

Considering all of this, we reject the assumption that participants in the “adaptable” condition simply solved more tasks because they were more open to the task or the topic itself. Thus, the higher amount of solved microtasks in this condition provides evidence to support **H2**. The statements “The game elements led to more fun” (“adaptable”: $M=3.8$, $SD=1.4$, $Mdn=4$; “fixed”: $M=3$, $SD=1.8$, $Mdn=3$; “without”: $M=2.7$, $SD=1.7$, $Mdn=2$; Kruskal-Wallis test: $p=.13$) and “The game elements motivated me to solve tasks more efficiently” (“adaptable”: $M=3.6$, $SD=1.6$, $Mdn=4$; “fixed”: $M=2.8$, $SD=1.3$, $Mdn=3$; “without”: $M=2.5$, $SD=1.4$, $Mdn=2$; Kruskal-Wallis test: $p=.07$) provided further support for this as the “adaptable” condition had the highest values here.

The “top-down” condition, and the gamification configuration initially selected automatically and unchanged in “selective top-down”, did not motivate participants to solve more tasks in comparison to the “no gamification” baseline. This stands in contrast to our findings with *ExpenseControl* and other gamification approaches, where it led to an increased performance (see Section 1.2.1 and Section 3.2.2). Statements regarding the meaningfulness of gamification in this setting, and that game elements motivated them, received mixed answers (with a mean between 3 or 4 on a 6-point scale). In general, this hints that for this population and this microtask setup, the game elements were potentially not suitable, which might explain why the performance in the “top-down”/“fixed” conditions was not better than in the “no gamification”/“without” condition. We can thus only partially (as “adaptable” was better than “without”) confirm **H1**.

Discussion

The goal of this study was to investigate whether “bottom-up” gamification is able to compete with or exceed “top-down” gamification. After re-grouping conditions the participants were in, we were able to identify differences (as hypothesized) in favor of “bottom-up” gamification. Participants that had a choice in how they wanted to use gamification, and used it, solved significantly more tasks faster without a decrease in correctness. We also found indications that they were not simply *more engaged overall* and therefore also tried out different game elements, but instead that “bottom-up” gamification was most likely the explanation for why they performed better. An important question is why 62% of the participants did not use the offered choices. Explanations for this could be that participants were not motivated by game elements in general, did not understand or comprehend their offered options, or did not see why they should use game elements at all. For this, the framing of the study (assisting researchers by doing simple tasks without any reward) might have been an explanation. As participants expected to remain on this page for only a short time, they may not have seen why they should also spend time on a peripheral feature, and instead simply focused on the task solving. Another explanation might be that certain users are not attracted by creating their own gamification setup (e.g., as they might dislike the additional effort) and simply did not want to try it. It was also unexpected to find no difference between the “top-down” gamification setting and the setting without gamification. We see two explanations for this: first, participants saw no benefit. In *ExpenseControl* the microtasks directly improved the algorithm for their digital household accounting book and they could also see how their receipt parts got corrected over time. Second, the framing of the study could have had an effect. Participants did not resume the work later. The household accounting book, on the other hand, was used several times during a week and differences in the game-element perception might become visible only after a longer interaction time. Overall, this might hint that in unrewarding scenarios, “bottom-up” might be more attractive than “top-down” gamification.

Our study had limitations: first, the number of participants, especially in the presence of five conditions, can be seen as a limiting factor. Second, the amount of options in the “selective” and “bottom-up” conditions made the study design rather complex, but was also a direct consequence of our previous study (see Section 4.3.2) that showed that a broad range of different game elements should be offered to account for every participant. Reducing the game elements to only a few might result in a situation in which some participants, even when giving a choice, could not find a setup that was appealing for them, which then would again confound the results. To reduce this complexity, follow-up studies could for example pre-select game elements fitting every player type [306] and present only these to the corresponding participants to reduce the complexity. Another option would then be to do a within-subject design in which all combinations of the pre-selected game elements are tested, and in one further condition provide participants with a choice to set up the combination on their own. Third, the lack of social connections (by having no friends in the system) is an issue for people that are motivated by, for example, competitive elements.

4.4.3 Contribution to the Thesis’ Questions

With the presented system (**Goal_{BU}Micro 1**), we were able to show that when people have influence options on the gamification aspects of a systems and use the offered choices, positive effects occur from a quantitatively measurable point of view (**Goal_{BU}Micro 2**). This further adds to **RQ2** (see Section 1.4). Even though the scenario we have chosen consisted of boring tasks (as reported by participants) and the solving had no further benefit to the participants (compared to *ExpenseControl*), the “bottom-up” approach was able to motivate users that adapted the gamification to solve more tasks. An interesting aspect was that the “top-down” approach that worked in the self-sustaining system *ExpenseControl*, did not spark enough interest here. For the thesis this is of particular interest: first, it shows that the idea of a self-sustaining system is appealing by itself (further adding to **RQ1**; see Section 1.4). As participants in this scenario here got no personal benefit from solving these tasks (i.e., this scenario did not represent a self-sustaining but only a crowdsourcing system), they did not engage much in it, even though it would have been possible to interact with the platforms over several days, similar to *ExpenseControl*. Second, the data suggest that a “bottom-up” approach could motivate participants more than a “top-down” setting in the context of unrewarding tasks, also hinting that user influence might in general be important. Both results need to be considered in respect to the stated limitations.

So far, we have seen that there are positive qualitative and quantitative effects when users receive influence on the motivational aspects of a system, but the previous studies have not considered why these are induced. It is currently unclear whether this difference can be accounted for by the choices offered (as described in different contexts in Section 2.2.3), by users being able to identify with *their* work [85, 237], or by the fact that participants selected a motivating

and suitable game configuration for themselves. The latter would mean that if participants were provided with a perfectly suited configuration in a “top-down” manner, this would have also led to positive results (and is the direction the personalization approaches take currently; see Section 2.3.1). All explanations support the idea of “bottom-up” gamification. If the choice itself or the self-identification aspect is what motivates participants, then providing choices for selecting which game elements should be used in gamified systems is reasonable. If, in contrast, the self-selected gamification configuration mattered, then again, offering participants a choice to select from a set of configurations or build their own configuration (as can both be done in “bottom-up” gamification) seems reasonable as well. The question of where the positive effects of “bottom-up” originate from is the driver for our next study.

4.5 Can Users Create Suitable Gamification Concepts?

Considering the aforementioned aspects, we wanted to investigate whether users are able to create suitable gamification configurations in terms of their personality and player types. To do this, the presented prototypes did not seem reasonable, based on their limited amount of offered game elements. In consequence, we had the following goals to receive further insights on “bottom-up” gamification attributes (e.g., which aspects should be empowered in such systems):

Goal_{BUSource} *Investigation of user-developed gamification concepts:* So far, we restricted users to a set of game elements to select from. Considering the “bottom-up” definition, this limits their choices. We want to explore what changes if users are not limited and are able to suggest gamification concepts as they see fit. As personality traits and player types have an impact on the perception of gamification (see Section 2.3), we were interested in whether this also transfers to self-developed gamification concepts. Consequently, when users select suitable game elements (in relation to their personality or player types) in a “bottom-up” scenario, this would be one factor for why “bottom-up” gamification induces positive effects.

4.5.1 User Study

We conducted an online study in which participants were asked to describe a gamification concept that would motivate them, for example, to do tasks more often. Following **Goal_{BUSource}**, we did not restrict participants to a pre-defined set of game elements from which they could select. Instead, we allowed them to describe what they wanted and saw fit to use. The resulting concepts were qualitatively evaluated for which game elements and mechanics were proposed. Basically, we had the following hypotheses:

Abbr.	Scenario
<i>En</i>	Imagine you want to save energy at your workplace, e.g., by turning off the lights after work. You have been tasked with developing a concept which has the goal to motivate you to save energy there (or motivate you even more, if you are already motivated).
<i>Ex</i>	You have been tasked with developing a concept which has the goal to motivate you to exercise more often, for example to go for a run multiple times a week.
<i>Pi</i>	Imagine you work for a manufacturer and build furniture by piece work (<i>"Piece work (or piecework) is any type of employment in which a worker is paid a fixed piece rate for each unit produced or action performed regardless of time"</i> (with a link to Wikipedia explaining this concept)). The work is monotonous and you have been tasked with developing a concept which has the goal to motivate you to do this job more thoroughly/more enjoyably/faster.
<i>Cl</i>	You have been tasked with developing a concept which has the goal to motivate you to clean the kitchen more often and faster.

Table 4.8: The scenario used in the gamification concept design study.

- H1** Participants can create gamification concepts when not restricted to a pre-defined set of game elements.
- H2** Participants select game elements that fit their personality in their self-created gamification concepts.

H1 was based on the previously seen study results in this chapter, in which participants were able to build motivational concepts from a pre-defined set of game elements. We hypothesized that even more options would lead to better fitting elements and that participants would still not be overwhelmed by the range of possibilities. **H2** was based on the broad range of literature (see Section 2.3) showing that player types or personality traits can be related to particular motivating game elements for users. A difference from our study is that these assessments are typically based on presenting the game elements one by one (e.g., [227, 306]) and participants are asked to rate them. In contrast, we will investigate what happens when users are asked to create motivational concepts. Based on this, if users are able to select "suitable game elements that motivate them", they should suggest those that are predicted by their personality.

Method

We set up four different online surveys in English. In each survey we stated that it was aimed at players of board/video games and that the goal was to analyze how these *solve a specific problem*. On the following page, participants were presented with five statements regarding their gaming affinity (e.g., *"I would characterize myself as gaming-affine"*). These statements were to be answered on a 5-point scale with the labels *disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, *agree*. Then, one of four scenarios (being the only difference between the surveys) was presented (see Table 4.8) with the task to develop a concept that would motivate them in this scenario. The scenarios were basically the same as those

we used in Section 4.2.2, in which participants had to state which game elements from a fixed set would motivate them in the scenarios. To reduce the workload, and based on the finding there that individuals do not differ across scenarios in their game element selection, participants needed to work on only one scenario. We required participants to write at least 700 characters, but also emphasized that writing more is appreciated, and we provided a “hook” (i.e., some aspects that might be considered in the concepts) as an initial starting point for their concepts [141]: *“This concept should be a game or game-like. It is up to you whether this is a digital (PC game, mobile game, Facebook game, ...) or analogue game (board game, card game, ...). You may decide freely among all aspects, for instance, whether this game is “just for you”, a game with friends or players you don’t know, etc. How would this game look?”*. This was followed by two additional free-text questions (*“Why do you think that your game concept will motivate you?”* and *“Which element of your concept is most important for you?”*). Seven statements (on the same scale as above) assessed the participant’s perception of the concept and the given scenario.

Afterward, the *Gamification User Types Hexad Scale* [306] and the *Big Five Inventory* with eleven items as suggested by Rammstedt and John [242] needed to be filled out (see Section 2.3). To inspect the relationships between the player types or the personality traits (which will be abbreviated as PTPT subsequently) and the game elements suggested in the concepts, we correlated the element usage in the concepts with the results of these scales. In addition, to learn about the characteristics of our sample, we followed the common way in which these relationships are investigated in the related literature (e.g., [227, 306]) and additionally presented twelve game element statements (which will be abbreviated as GES subsequently), e.g., *“Please state for every game element how motivating it is for you in general: Being able to unlock new features and/or content in a game”*, that needed to be rated on a 4-point or 5-point scale (*Not motivating, Somewhat motivating, Moderately motivating, Very motivating, Extremely motivating*)³⁷. As the Hexad scale is relatively new, we based the GES on Tondello et al.’s research [306] and used their principal elements (or design elements in the case of the *Philanthropist*), to be able to compare the results (using two GES per player type). Following H2, we expect that answers to the GES correlate with the PTPT (as shown by Tondello et al.). We hypothesized that there is a connection between answers to the GES and the game elements suggested in the concepts, as both originated from the question *“What would motivate you?”*. As we could not anticipate, at the questionnaire setup, which game elements would actually be suggested in the concepts, it was to be expected that not every GES eventually would have a related code (e.g., for the GES example, the code to be found would be Unlockables in the concepts).

We used *Amazon Mechanical Turk* (AMT) (see Section 2.1) to distribute the four questionnaires, restricted the selection to US Turkers and ensured that every Turker could only participate in one questionnaire. We added four test questions (in which they were asked to select a particular answer) to check that participants

³⁷ We started with a 4-point scale for *En* but changed this to a 5-point one for the other scenarios to give better discriminative options.

did not simply rush through the questions [197]. We pre-tested the questionnaire with twelve students and university employees, and on AMT with ten participants, to see how long it takes to fill out the questionnaire, to learn about potential issues and to receive a first set of answers to develop an initial code book (see below). On average, filling out the questionnaire took 15 minutes in the pre-test. Thus, we paid \$1.50 to meet the minimum wage suggestions of [259].

Coding process

We analyzed the written concepts to learn which game aspects were mentioned (which will be further described as “elements”, independent of their abstraction level [64, 92]), by conducting a content analysis [118]. All coders were gamification researchers: two independent coders inspected the twelve answers from the first pre-test separately to develop a first code book version. For the code book development, we followed an inclusive approach that did not consider only typical gamification elements, i.e., for example, Themed or Communication Tools were also codes that were added. The first code book was used by these two and one additional coder to code the remaining ten answers from the second pre-test. The coding results were discussed, deviations solved and the code book refined accordingly. The coders reported that no situation occurred where they would have liked to discuss something with the participants, indicating that a survey as an instrument appears suitable. Certain codes had relationships, e.g., if Social Competition was coded, Social Comparison was coded as well. With the resulting set of codes (see Table 4.9), the 140 answers of the main study were coded by two coders independently. Afterward, they went through their results and solved deviations via discussion. To check the validity of this coding, a third coder coded a random sample of 42 participants (30%) and the inter-rater agreement for every code was calculated, which was on average $\kappa = .86$ (Min=.63; Max=1). This can be considered as “almost perfect” [283].

Results

We used AMT until we received 35 valid responses³⁸ for every scenario, summing up to 140 participants (79 male, 60 female, one no answer; age: <18: 1, 18–24: 5, 25–31: 41, 32–38: 50, 39–45: 17, 46–52: 12, 53–59: 10, >59: 3, no answer: 1). Mainly, participants were employed for wages (69%). From the most often selected educational levels, 36 (26%) had some college credit and 50 (36%) had a bachelor’s degree. All participants can be considered as open to games, as they at least answered with a neutral response to one of our gaming related questions (“Do you characterize yourself as gaming-affine?” / “Do you frequently play video (board) games?” / “Do you have a passion for video (board) games?”). 32 participants (23%) also stated that they had designed a game already; these were distributed across

³⁸ This number was an economical decision. We judged this number as acceptable compromise for being able to code in a reasonable time frame and still offer enough expressiveness.

the four scenarios (10/6/6/10, denoting *En/Ex/Pi/Cl*), so that we did not expect issues here. The average time spent for filling out the different questionnaires was higher than in the pre-tests (18.6 min/17.1 min/18.5 min/17.3 min) but did not differ significantly across the scenarios as a Kruskal-Wallis test showed ($p=.74$).

Diverse gamification concepts and game elements were suggested: We analyzed the average character count of the created concepts in the scenarios (1102/939/1072/925; Kruskal-Wallis test $p=.11$), and between participants claiming to have developed a game ($M=1112$) and those who had not ($M=973$), but no significant effects were found (t -test $p=.14$). By inspecting the concepts, we found that participants suggested diverse ones, based on the codes identified. The following examples³⁹ are taken from *Ex* and are shown with the codes found (see Table 4.9 for further explanations to these codes):

“The player would be rewarded for exercising by getting points they can accumulate and add up during the day. These determine how many cards they can pick from the reward deck and which deck they are allowed to choose. The cards in each deck will be written by the player and could be anything they would enjoy doing. Let’s say you exercise for 30 minutes; you could either pick two cards from the first or one from the second deck. If they exercise for one hour the cards double. The rewards in the second deck will be more exciting, but you only get one. Deck one might have a card that says “social media access for 20 minutes, 30 minutes TV viewing, 20 minutes of music videos”. Deck two will have more exciting stuff like “video gaming for one hour, Netflix for one hour, etc.”. The game is a board game style meant for one player and personalized by the players themselves.”

Codes: Goals, Periodicity, Single-Player, Points, Bonus, Prizes, Collecting/Collectibles, Customization, Surprise, Board Game

“The game that got me outside the fastest and the most was Pokémon Go. There could be some real changes that would get me out and running. However, it’d be difficult to come up with a knockoff that didn’t feel like a knockoff. Maybe something with a similar theme (augmented reality) but different content (Pokémon). Something educational might be neat. There was a step app I used a while back, I can’t remember the name, but it would count my steps along a path to somewhere (like a trip around Rome). Every x00 steps I would get a notification that I had reached a landmark and I could click on it and view the information about the landmark on my trip. Maybe you could build something in Google Maps/Street View so that one could walk around and look at the phone and virtually walk around a city in another country.”

Codes: Visible Progress, Progression, Surveillance, System Assistance, Notification, Knowledge/Skill Improvement, Exploration, Mobile Game, Analogy

³⁹ Shortened and grammatically corrected for presentation reasons.

Rank	Game Element	Explanation – The participant mentions ...	Freq.
01	Goals	... specific (sub-) goals that need to/can be achieved	106 (76%)
02	Multiplayer	... a multiplayer component	87 (62%)
03	Progression	... progression in the game or specific game attributes	81 (58%)
04	Social Comparison	... that it is possible to compare one's own performance to others'	72 (51%)
05	Social Competition	... competition between human players where one player will be the winner	70 (50%)
06	Points	... an entity that can be accumulated in the game	66 (47%)
07	Prizes	... a physical (e.g., cinema tickets) or a personal reward (e.g., to watch TV)	65 (46%)
08	Mobile Game	... a mobile component of the game	63 (45%)
09	Time Pressure	... actions that need to be done within a certain time	49 (35%)
10	Periodicity	... something that is or should be done regularly	42 (30%)
11	Visible Progress	... an indication that shows progress or distance to the next goal	40 (29%)
12	Bonus	... that it is possible to achieve a bonus (e.g., receiving 1000 bonus points)	35 (25%)
13	Surprise	... randomness or other unexpected elements that surprise the player in the game	34 (24%)
14	Surveillance	... a control instance that monitors progress in the game	33 (24%)
15	Friends	... that the game can be played with friends/family	28 (20%)
16	Customization	... that the player is able to adjust components of the game to his or her needs	28 (20%)
17	Achievements	... achievements (e.g., badges, ranks) can be gained	27 (19%)
18	Unlockables	... that new game content or features can be unlocked	26 (19%)
19	Analogy	... an existing game or concept that he or she adapts to his or her concept	26 (19%)
20	Punishment	... penalties (e.g., subtracting points)	24 (17%)
21	Social Recognition	... that progress made is visible to others, with the purpose to show it to them	21 (15%)
22	System Assistance	... an assistance function of the game that eases something for the player	20 (14%)
23	Virtual Self	... one or more virtual characters that represent the player	20 (14%)
24	Real Challenge	... a particularly challenging aspect or beating one's own personal scores	17 (12%)
25	Themed	... that the concept has a specific theme (e.g., Sci-Fi)	17 (12%)
26	Single-Player	... that the concept is (also) usable for a single player	16 (11%)
27	Virtual Items	... virtual items/goods that are available in the game	15 (11%)
28	Knowledge/Skill Improv.	... that the game (or players) conveys knowledge (to others) or that players can improve their skills	15 (11%)
29	Teams	... that players can be grouped into teams or guilds	14 (10%)
30	Notification	... that the game provides a notification when something particular happens	13 (9%)
31	Unspecific Reward	... rewards, but what kind is not further specified	13 (9%)

Table 4.9: The game elements found in the gamification concept design study, an explanation and their frequency.

Rank	Game Element	Explanation – The participant mentions ...	Freq.
32	Appearance	... something that is part of to the look/appearance/sound of the game	12 (9%)
33	Board Game	... that the game is a card or board game	12 (9%)
34	Encouragement	... that the game or other players encourage one to reach goals	11 (8%)
35	Social Collaboration	... that players cooperate to reach a goal	10 (7%)
36	Collecting/Collectibles	... collecting a specific entity explicitly as a motivational factor	10 (7%)
37	Anti-Cheating	... something that prevents cheating in the game	9 (6%)
38	Unspecific Competition	... a competition, whether it is against other players or computer-controlled entities	8 (6%)
39	Exploration	... that features in the game, the game itself or the world can be explored	8 (6%)
40	Lottery/Gambling	... that players can bet in the game or have a raffle	8 (6%)
41	RPG	... role-playing games explicitly	8 (6%)
42	Socialization	... that the game helps to get to know other people (or improve a relationship)	7 (5%)
43	Fairness	... concepts that make the game fair (e.g., goals depending on the fitness level)	7 (5%)
44	Peer Pressure/Accountability	... that other people can see whether or how I do or fail to do something	6 (4%)
45	Personalization	... that the system adapts itself to the player's needs	6 (4%)
46	Virtual Character (other)	... one or more characters that represent not the player, but other entities (e.g., NPCs)	6 (4%)
47	Common Welfare	... that the game serves a higher purpose (e.g., saving the world)	6 (4%)
48	Creativity	... that the player needs to be creative (e.g., showing a scene from a movie)	6 (4%)
49	Puzzle	... a riddle or puzzle component	5 (4%)
50	Communication	... that it is possible to communicate with others	4 (3%)
51	Security/Privacy	... security or privacy aspects	3 (2%)
52	Care Taking	... that the player needs to care for others (e.g., virtual pets)	3 (2%)
53	Premium/Freemium	... special features that can be purchased, or the availability of specific features for free	3 (2%)
54	Mini Games	... mini games that can be played within the game	3 (2%)
55	Persuading/Manipulating	... specific actions that can persuade/manipulate others so that they adapt their behaviors	2 (1%)
56	Story	... story components	2 (1%)

Table 4.9: The game elements found in the gamification concept design study, an explanation and their frequency (cont.).

Overall, the 140 participants suggested 1348 elements in their concepts, with 9.6 elements on average ($SD=3.2$, $Min=3$, $Max=19$), with no significant difference between the scenarios (9.9/10.3/9.1/9.1; Kruskal-Wallis test, $p=.36$). Participants reported being satisfied with their concepts ($M=4.5$, $SD=.6$, $Mdn=5$), that they were easy to develop ($M=4.2$, $SD=.9$, $Mdn=4$) and that they would be motivating for themselves ($M=4.6$, $SD=.7$, $Mdn=5$) and others ($M=4.6$, $SD=.5$, $Mdn=5$). The participants state that they could imagine the scenarios ($M=4.4$, $SD=.8$, $Mdn=5$) and that these were relevant for them ($M=4.2$, $SD=1$, $Mdn=5$). We analyzed whether the number of game elements suggested differ between those who had or had not designed a game already; the responses to the five gaming affinity questions; and the answers to the concept/scenario questions above, but no significant differences were found (always $p>.05$).

We inspected all concepts and found that no set of game elements was proposed twice, i.e., participants provided concepts that used different elements. When relaxing this to sets which deviate by just one element, we still found no overlap (two different elements: 4 sets (3%), three: 16 (11%), four: 36 (26%)). Table 4.9 reveals that only five of the 56 elements (9%) were mentioned by at least 50% of our participants. This shows that in an unrestricted design task, many different elements seem to be of relevance for participants. By considering the ten most often mentioned elements, we see that many participants defined a Goal (rank 01) and Progression (03) in their concepts. Interestingly, even though we biased (through the formulation of the “hook”) Single-Player (26) and Multiplayer (02) games, the latter appeared to be more relevant for participants. If the game had a multiplayer component, we also learned that most participants suggested a competitive element (04, 05) instead of a collaborative one (35). A Mobile Game (08) was also much more frequently selected than, for example, a Board Game (33), although both were part of the “hook”. Points (06) and Prizes (07) were the most often mentioned reward types and many participants considered a timing component (09, 10). For these aspects, it needs to be kept in mind that the ranks beginning with 06 were mentioned by less than 50% of our participants. In general, these aspects also show the participants’ diversity when the system does not restrict them. Overall, the number of elements suggested, the variety of the concepts and the self-reports provided support for **H1**.

We found 23 significantly different game element usages in the concepts between the scenarios (see Table 4.10). While Time Pressure seems to be explainable by the scenario framing (as in *Cl* and *Pi*, doing it faster was highlighted), the other differences appear to arise from the scenario itself. In *Ex*, participants seem to want systems that monitor what they (or others) do and they want to have the option to customize systems to their needs. In *En* collaborative aspects were dominant, and a higher “purpose” was highlighted. In *Cl*, friends/family were more often mentioned, most likely because the kitchen area is a private room where friends and family might have access. Finally, *Pi* revealed that participants want to use more personal challenges here. This shows that the context had a moderating effect on which game elements were more likely to be suggested.

Game Element	En	Ex	Cl	Pi
Time Pressure	1	9	21 ^{Ex, En}	18 ^{En}
Surveillance	9	14 ^{Pi}	7	3
Friends	3	7	14 ^{En, Pi}	4
Customization	4	14 ^{En, Pi}	7	3
Achievements	3	13 ^{En, Pi}	8	3
Analogy	4	12 ^{Cl}	3	7
Real Challenge	2	2	3	10 ^{Ex, En}
Teams	9 ^{Ex, Cl, Pi}	2	1	2
Notification	8 ^{Pi}	3	2	0
Social Collaboration	7 ^{Ex, Cl}	1	0	2
Exploration	1	6 ^{Cl}	0	1
Purpose/Common Welfare	6 ^{Ex, Cl, Pi}	0	0	0

Table 4.10: Significantly different element suggestions (Kruskal-Wallis tests with pairwise comparisons (Bonferroni-corrected), all $p < .05$) between scenarios in the gamification concept design study. Superscripts denote pairwise-comparison results (for readability only mentioned once per relationship).

PTPT seem not to be a dominant factor in this task: In the following, we will present several complex tables. Before we derive the corresponding results from these, we will give an explanation on how to read and interpret the data.

How to read Table 4.11: We will provide an ongoing example to illustrate how to read the tables throughout this section. Nine GES also had a related element (abbreviated with RE subsequently) in the concepts, i.e., we found a game element in the concepts that also was represented in the GES. We had twelve GES, but only found corresponding elements for nine (with two GES having two REs); only these are presented and shown in the first column of Table 4.11. For the ongoing example, we will always consider the first row of the corresponding table: the GES here is “Being confronted with challenges that push me to my limits” with the RE Real Challenge. Based on the previous result that the scenarios appeared to have an influence, the subsequent considerations are done scenario-wise. The second column thus indicates the scenario and the number of participants who used the RE here. For our example, in *En* and *Ex* two, in *Cl* three and in *Pi* ten participants used the game element Real Challenge in their gamification concepts. The third column contains the mean values to the answers to the GES per scenario (5-point scale for *Ex*, *Cl*, *Pi* and 4-point scale for *En*). 35 participants per scenario answered the GES. We differentiate the means of those who did not use the corresponding RE in their concepts ($M_{!Used}$) from those who did use it (M_{Used}). We also show the p-value of the Mann-Whitney U test where we compared $M_{!Used}$ with M_{Used} . Considering our example for *Pi* only, we see that the ten participants (of the 35 participants per scenario) who used the element Real Challenge in their concepts had answered the GES with a mean value of 4.4 on the 5-point scale, while the remaining 25 participants, who did not use Real Challenge in their concept, had a mean value of 3.8. The p-value for the comparison of both means was $p = .05$.

Game element statement (GES) and related element (RE)	Scenario and # of RE mentions		Answers to GES		
			M_{Used}	M_{Used}	p
<i>Being confronted with challenges that push me to my limits</i> Real Challenge	En	2	3.3	3.5	.76
	Ex	2	4.1	3.5	.34
	Cl	3	4	4.3	.72
	Pi	10	3.8	4.4	.05
<i>Having quests/tasks/ missions I can solve</i> Goals	En	21	3.7	3.8	.65
	Ex	28	4.6	4.5	.95
	Cl	28	4.6	4.3	.67
	Pi	29	3.8	4.4	.12
<i>Being able to unlock new features and/or content in a game</i> Unlockables	En	6	3.7	3.3	.31
	Ex	12	3.2	4.5	.33
	Cl	4	4	4.8	.18
	Pi	4	4.3	4	.53
<i>Having tasks that allow me to explore aspects and features in a game</i> Exploration	En	1	3.6	3	.46
	Ex	6	4.1	3.7	.31
	Cl	0	3.9	-	-
	Pi	1	4	2	.11
<i>Having the possibility to share my knowledge with others</i> Knowledge/Skill Im.	En	6	3	3.3	.38
	Ex	3	3.2	4	.26
	Cl	3	3.2	3	.72
	Pi	3	3.4	2.3	.16
<i>Receiving badges/ achievements</i> Achievements	En	3	3.4	3.7	.56
	Ex	13	3.6	4.4	.11
	Cl	8	3.5	4.3	.05
	Pi	3	3.7	4.7	.16
<i>Receiving points. I can compare with others on a leaderboard</i> Points	En	21	3.1	3.5	.14
	Ex	12	3.9	4.1	.88
	Cl	18	3.8	3.7	.96
	Pi	15	3.9	3.7	.63
<i>Receiving points. I can compare with others on a leaderboard</i> Social Competition	En	17	3.2	3.5	.25
	Ex	13	3.9	4	.91
	Cl	20	3.5	4.1	.30
	Pi	20	3.7	3.9	.66
<i>Having the option to build guilds/teams to solve tasks together</i> Teams	En	9	2.9	3.1	.67
	Ex	2	3.4	2	.27
	Cl	1	2.7	3	.80
	Pi	2	3.1	3	.97
<i>Having the option to build guilds/teams to solve tasks together</i> Social Collaboration	En	7	2.9	3	.95
	Ex	1	3.3	1	.17
	Cl	0	2.7	-	-
	Pi	2	3.1	3	.97
<i>Having features that help me to get to know other people</i> Socialization	En	1	2.7	4	.17
	Ex	3	2.7	3.7	.34
	Cl	2	2.8	2.5	.81
	Pi	1	2.7	3	.80

Table 4.11: Scenario-wise comparison (Mann-Whitney U) of the GES mean answers of those who did and who did not use the element in their concept.

Result interpretation of Table 4.11: We expected relationships between the GES and the RE as both were asked in the context of “What motivates you?”. Considering Table 4.11, though, we found no significant differences between the mean of the GES answers of those who did or did not suggest the element, i.e., participants who suggested it did not provide a higher rating for the corresponding GES. In the 366 cases in which participants suggested a RE, 64 corresponding GES ratings (17%) were below 4 on the 5-point scale (3 for *En* on the 4-point scale), i.e., they did not rate the element as particularly motivating but used it in their concepts. In the 1174 cases in which participants did not suggest a RE, 334 GES ratings (28%) were above 3 (2 for *En*), i.e., the element was rated as motivational but was not used in the concept. Although hypothesized differently, both hint that there is no clear relationship.

How to read Table 4.12 and Table 4.13: In both tables we repeat the GES and the RE and report the RE per scenario in the first column. All following columns show the *Hexad* user types (for Table 4.12) correlated (Kendall’s τ) with the answers to the GES and the usage of the RE in the concepts. A cell entry is only shown if at least one of these two revealed a significant correlation at the $p < .05$ level. If a significant correlation between a player type and the GES exists, the value of τ is shown in bold to the left of /. If a significant correlation between player type and the RE exists, the value of τ is shown in bold to the right of /. Furthermore, significant correlations are colored green (red) if the correlation is positive (negative). The cell is highlighted in blue if both correlations are significant and have the same direction. If only one significant correlation is found, we always show τ for the other non-significant correlation (non-bold and in black). Considering the ongoing example, we found five significant correlations for *Pi* and the player types: four between the player types and the GES (*Achiever*, *Free Spirit*, *Philanthropist*, *Player*; left of the /) and one for the player types and the RE (*Achiever*; right of the /). These were all positive (highlighted in green). The *Achiever* cell is highlighted, as the GES with the *Achiever* and the RE with the *Achiever* both correlated significantly and had the same direction. As a significant correlation between *Free Spirit*, *Philanthropist* and *Player* and the RE was not found, but significant ones were found with the GES, we denote their τ value to the right of the / (non-bold and written in black). For the player types *Disruptor* and *Socialiser*, no significant correlations at the $p < .05$ level were found for either the GES or the RE and player types. Thus, the cells are empty.

Table 4.13 can be read analogously, but here the *Big Five* instead of the player types are considered. For our example, three significant correlations for *Pi* were found: one for the *Big Five* and the GES (*Conscientiousness*, left of the /) and two for the *Big Five* and the RE (*Conscientiousness* and *Agreeableness*, right of the /). The cell for *Conscientiousness* is highlighted, as both correlations are significant and had the same direction. For *Openness*, *Neuroticism* and *Extraversion* no significant correlations were found and therefore the cells are empty.

Game element stat. (GES) and related element (RE)	Correlations: Hexad user types and [GES/RE]					
	Achiev.	Disrupt.	Free Sp.	Philan.	Player	Sociali.
<i>Being confronted with (En 2) challenges that push (Ex 2) me to my limits (CI 3)</i>			.30 / -.15			
Real Challenge (Pi 10)	.55 / .31		.53 / .22	.42 / .19	.47 / .06	
<i>Having quests/tasks/ (En 21) missions I can solve (Ex 28) (CI 28)</i>	.41 / .33		.39 / -.04	.44 / .05	.58 / -.05	.30 / -.06
Goals (Pi 29)	.38 / .30		.30 / .37	.40 / .31	.47 / .29	
<i>Being able to unlock (En 6) new features and/or (Ex 12) content in a game (CI 4)</i>	.32 / -.15		.30 / -.06		.36 / -.05	
Unlockables (Pi 4)	.34 / .02			.34 / -.22	.39 / -.06	.10 / -.35
<i>Having tasks that allow (En 1) me to explore aspects (Ex 6) and features in a game (CI 0)</i>	.51 / -.13		.39 / -.06	.57 / -.14	.48 / -.27	
Exploration (Pi 1)						
<i>Having the possibility (En 6) to share my knowledge (Ex 3) with others (CI 3)</i>	.32 / .10		.28 / -	.44 / .05	.28 / .05	.30 / -.13
Knowledge/Skill Im. (Pi 3)	.52 / -.25		.37 / -.25	.53 / -.24	.45 / -.21	.38 / -.13
<i>Receiving badges/ (En 3) achievements (Ex 13) (CI 8)</i>		-.40 / .09				
Achievements (Pi 3)	.34 / -.10		-.02 / -.35	.33 / -.04	.30 / -.01	
<i>Receiving points. (En 21) I can compare with (Ex 12) others on a leaderboard (CI 18)</i>		-.38 / -.17		.47 / .09	.27 / .04	
Points (Pi 15)	.37 / -.07			.33 / .09	.29 / -.05	.41 / .14
<i>Receiving points. (En 17) I can compare with (Ex 13) others on a leaderboard (CI 20)</i>	.41 / .07			.30 / .07	.34 / -.18	
Social Competition (Pi 20)	.34 / -.09			.31 / -.32	.41 / .09	
<i>Having the option to (En 9) build guilds/teams to (Ex 2) solve tasks together (CI 1)</i>					.41 / -.07	
Teams (Pi 2)	.37 / -.11			.33 / -.28	.29 / -.10	.41 / -.10
<i>Having the option to (En 7) build guilds/teams to (Ex 1) solve tasks together (CI 0)</i>	.41 / .15			.30 / -.11	.34 / .00	
Social Collaboration (Pi 2)	.34 / .03			.31 / -.19	.41 / -.10	
<i>Having features that (En 1) help me to get to (Ex 3) know other people (CI 2)</i>						.34 / .11
Socialization (Pi 1)						.43 / -.10
						.57 / .07
						.26 / -.11
						.34 / .00
						.43 / -.23
						.57 / -
						.26 / -.11
						.49 / .24
	.27 / .04			.40 / .14		.45 / .25
						.35 / .03

Table 4.12: Correlations (Kendall's τ , colored if $p < .05$) between GES and RE with Hexad user types (cell empty if both correlations are $p > .05$ or highlighted when in the same direction).

Game element stat. (GES) and related element (RE)	Correlations: Big Five personality traits and [GES/RE]				
	Open.	Consc.	Neuro.	Extrav.	Agree.
<i>Being confronted with</i> (En 2)	.36 /.10		-.35 /.00		
<i>challenges that push</i> (Ex 2)	.32 /-.21	.36 /.02		.29 /-.10	
<i>me to my limits</i> (CI 3)					
Real Challenge (Pi 10)		.33 /.40			.05/ .30
<i>Having quests/tasks/</i> (En 21)	.42 /.17		.17/ .39		
<i>missions I can solve</i> (Ex 28)	.40 /-.21	.49 /.02			
(CI 28)					
Goals (Pi 29)		.30 /.36			
<i>Being able to unlock</i> (En 6)	.46 /-.25	.43 /-.08			
<i>new features and/or</i> (Ex 12)		.50 /-.08		-.01/ -.30	
<i>content in a game</i> (CI 4)			.30 /.26		
Unlockables (Pi 4)					.42 /-.10
<i>Having tasks that allow</i> (En 1)	.48 /-.07		-.31 /-.05	.32 /.16	.42 /.07
<i>me to explore aspects</i> (Ex 6)		.32 /.00			
<i>and features in a game</i> (CI 0)					
Exploration (Pi 1)					.32 /-.15
<i>Having the possibility</i> (En 6)				.51 /.11	.41 /.23
<i>to share my knowledge</i> (Ex 3)					
<i>with others</i> (CI 3)	.10/ -.35		-.29 /.20	.37 /.02	.02/.00
Knowledge/Skill Im. (Pi 3)				.29 /-.20	.36 /-.28
<i>Receiving badges/</i> (En 3)				-.01/ .33	
<i>achievements</i> (Ex 13)					
(CI 8)					
Achievements (Pi 3)					.29 /.12
<i>Receiving points.</i> (En 21)			-.31 /.01		
<i>I can compare with</i> (Ex 12)	.24/ -.40	.43 /-.04		.30 /.06	
<i>others on a leaderboard</i> (CI 18)					.32 /.18
Points (Pi 15)					
<i>Receiving points.</i> (En 17)			-.31 /-.49		
<i>I can compare with</i> (Ex 13)		.43 /.07		.30 /-.04	
<i>others on a leaderboard</i> (CI 20)					.32 /.02
Social Competition (Pi 20)					
<i>Having the option to</i> (En 9)					
<i>build guilds/teams to</i> (Ex 2)					
<i>solve tasks together</i> (CI 1)				.44 /-.05	.25/-
Teams (Pi 2)			-.09/ .30	.29 /-.04	
<i>Having the option to</i> (En 7)					
<i>build guilds/teams to</i> (Ex 1)					
<i>solve tasks together</i> (CI 0)				.44 /-	.25/-
Social Collaboration (Pi 2)			-.09/ .30	.29 /-.04	
<i>Having features that</i> (En 1)				.30 /.25	
<i>help me to get to</i> (Ex 3)					
<i>know other people</i> (CI 2)				.47 /-.12	
Socialization (Pi 1)				.38 /.19	

Table 4.13: Correlations (Kendall's τ , colored if $p < .05$) between GES and RE with the Big Five (cell empty if both correlations are $p > .05$ or highlighted when in the same direction).

Game Element	Achiev.	Disrupt.	Free Sp.	Philan.	Player	Sociali.
(Real) Challenge	.46	.21	.41	.21	.32	-
Quests (Goals)	.27	-	.24	-	.25	-
Unlockables	-	-	.23	-	-	-
Exploratory Tasks (Exploration)	-	-	.35	-	-	-
Knowledge Sharing	-	-	-	-	.23	-
Achievements	.21	-	-	-	.27	-
Points	-	-	.20	-	.26	-
Social Competition	-	.32	.25	-	.24	.22
Teams	-	-	-	-	-	-
Social Networks (Socialization)	-	-	-	-	-	-

Table 4.14: Correlations between *Hexad* user types and elements found in [306] (excerpt). We only show a τ -value if $p < .01$ and $\tau > .2$.

Result interpretation of Table 4.12 and Table 4.13: By considering the two tables, as expected (based on the literature reporting correlations between game elements and PTPT [227, 306]), we found 145 significant correlations between the GES and PTPT in the different scenarios, indicating a general relationship. Although the GES were the same in every scenario, not all correlations were found in every scenario consistently: this only happened in 20 of the 145 cases, indicating that the scenarios might have a stronger effect. Comparing our correlations between GES and the player types with the correlations reported in [306], we see, although some overlap exists, that we differ in the size and amount of the correlations (see Table 4.14). One notable example here is that we found several relationships to the *Philanthropist*, while Tondello et al. reported only a few weak ones. We also considered the significant correlations between the PTPT and the RE, but only 22 were found, of which only nine are in line with the significant correlations found with the GES. One issue here is the low usage count of many elements in the concepts. But even when considering the 15 cases in which elements (i.e., Goals, Social Competition, Points, Achievements and Unlockables) were suggested ≥ 10 times per scenario, the number of the same correlations between PTPT and GES and PTPT and RE is low. This hints that neither player types nor personality traits seem to be a dominant factor for which elements are suggested in such an open design task. This suggests rejecting **H2**.

Discussion

Considering “bottom-up” gamification the study shows several aspects that add to **Goal_{BUSource}**: first, the study revealed that participants are able to describe gamification concepts without guidance in a task where they have a lot of freedom, i.e., participants seem not to be overwhelmed by this (adding to **H1**). Second, participants utilize a broad range of game elements in their concepts. As no set of elements was suggested twice, the game element overlap in the configurations is comparatively low. Only five elements were mentioned by at

least 50% of the participants, showing the diversity in this task. For “bottom-up” settings providing predefined set of elements, this means that a broad range should be offered (in particular, more than was offered in our previous studies) to account for the different user preferences. Our presented ranking of game elements can be used to guide the selection. Third, although originating from self-report data and thus needing to be treated with caution [235], participants claimed nearly uniformly to be satisfied with their concepts, that it was easy for them to develop these and that they think that these would be motivating for themselves and others. The latter was integrated to mitigate the effects of potential bias on the self-reports to a certain degree [3]. Although these concepts were not implemented and tested to validate the factual motivational impact, this at least hints that participants seem to be convinced of what they have created on their own (also adding to **H1**). Fourth, considering the study of *BU-ToDo* (see Section 4.3.2), the participants there actually only used a few game elements and combined these only into simple gamification concepts. But when asked, they stated they would have used more complex setups, but had no opportunities to do so because of the short study duration. As our participants here suggested many elements and complex concepts, this supports this finding further.

We found that some participants who suggested specific elements in their concepts (which had the goal to motivate them) did not rate them as particularly motivating in the GES (and vice versa). This either indicates that specific elements only become motivating for participants in combination with other elements (which would be an issue for related work that investigates game elements individually) or that participants develop concepts with elements that are not particularly motivating, but just, for example, *known* to them. As we have seen that the context (i.e., the scenario) also has an impact on the game element selection, this can be another explanation, as the GES questions had a different context (i.e., a general one). We were unable to replicate some of the relationships between player type and the GES reported in [306] and instead found different correlations. An explanation for this is that the formulations/explanations used for the game elements were different (as these were not provided in [306]) or that more research for the *Gamification User Types Hexad Scale* is necessary to come to a consistent relationship, e.g., to develop standardized formulations and rules (e.g., in which context these questions should be framed) for how to assess game elements. As the *Big Five* literature also provides contradicting results (e.g., [131] and [224] report contradicting relationships for the competition element), we have not further analyzed the relationships found between our GES and the *Big Five*. In general, our correlations with the GES were not stable across scenarios. One explanation for this might be that the scenario had a priming effect. Even though the GES were formulated without a particular context, participants (potentially subliminally) might still consider these for the scenario they had worked on before. As we found indications that the context affects the game element usage and was shown to be relevant for gamification in general [106], and as personality traits are also affected by context [201], this seems plausible.

Considering the PTPT and the elements suggested in the concepts, we found fewer correlations than found with the GES. One explanation is that some of the elements were suggested too infrequently in comparison to the GES being answered by every participant. But even those that were suggested often matched the GES only infrequently. This suggests that the PTPT seem not be a dominant factor in the element selection in such a design task. Thus, we tend to reject **H2**. Other factors seem to be more relevant here. One factor we have found in our study is (again) the context. This contrasts with the findings of the online study (see Section 4.2.2) where individuals stayed nearly consistent with their game selections across the four scenarios. Here, instead, we found evidence that the scenario impacts the game element selection. One explanation for this is that the options were different (free choice vs. selecting from a set). Another one is that the common game elements used in the online study in Section 4.2.2 are less likely affected by the context. Conducting further work on this seems relevant. Overall these findings suggest that it seems less likely that users are motivated in “bottom-up” settings because they selected suitable game elements for themselves, at least from a personality/player type perspective. Other explanations, such as “having a choice”, being able to identify with the self-created setup, or that certain elements might be easier to employ than others in such a task, might be more likely explanations. These are all in line with the thesis’ scope, i.e., that users should receive more influence options on systems.

This study has limitations: first, as the scenarios had a moderating impact, we considered them separately. This leads to fewer participants for the calculations. But even when considering the correlations of the PTPT and GES/RE of the whole sample, the results remain similar: 76 significant correlations between the GES and PTPT but only 14 between the RE and PTPT are found, of which only five are in line with the significant correlations found with the GES. Second, as no concept was implemented and used by the participant who developed it, we cannot derive whether the implemented gamification concepts would indeed be motivating for them, aside from the motivational questions. In our other reported studies in this chapter participants were able to realize their ideas without writing them down first, and positive results were found. Given that they mainly used their initial ideas throughout these studies, we assume that the written concepts would also induce positive effects when realized. Third, it also needs to be stated that participants in our study actually had a design task at design time. Although this aspect is also a part of “bottom-up” gamification (whenever the gamification concept is adjusted by participants), it is unclear how this relates to a real application. This was necessary to provide them with the freedom we wanted to achieve with **Goal_{BUSource}**. Fourth, in the context of self-reports, we also note that participants might have feared a rejection of their contribution in the *AMT* context (and thus not being paid) if they answered that their concept would not motivate them. We compared their answers to the “motivation” questions to our (unpaid) student pre-test sample and were not able to find a difference, but to rule this out, a larger unpaid sample needs to be tested.

4.5.2 Contribution to the Thesis' Questions

The study adds to our understanding of self-tailored gamification and thus to **RQ2** (see Section 1.4). Especially in relation to the limitations, it can only be seen as a first step. Such studies will help to understand what drives the positive “bottom-up” gamification effects. Here, we learned how users develop concepts when not restricted by an actual implementation or a pre-defined set of options in a questionnaire. Based on this study, apparently, the player types and the personality traits of the participants had no dominant impact on which game elements were suggested. Future work can build on this further and identify whether self-tailored gamification is something that can be used in general for all users, or is also only a facet working for a particular user type (see Section 2.3).

4.6 Summary

In this chapter, we considered self-tailored gamification by introducing and analyzing “bottom-up” gamification: the option of users to tailor the gamification in a system to their own needs at runtime. While customization approaches are not new (see Section 2.3.2), here the extent to which “bottom-up” gamification allows users to influence the system is. Users can decide whether to use gamification at all, can combine game elements as they see fit and can further customize them. We presented two systems, *BU-ToDo* and the *BU-Microtasks Platform*, that offered “bottom-up” gamification. These showed how such an approach can be implemented. Both systems were evaluated within user studies that provided positive results towards the systems themselves. But these user studies also validated the idea that providing users with more influence in a system is beneficial. Participants in the studies reported subjectively positive effects this kind of intervention has on them. Additionally, with a quantitatively measurable dependent variable we could show that people who could customize gamification, and did it, performed better than those who had no choice or had one and did not use it. Considering the ongoing efforts to move away from “one-size-fits-all” gamification approaches, and the knowledge that personalization approaches are not yet optimal (see Section 2.3 for both aspects), “bottom-up” gamification could thus be a valid alternative. We also have analyzed whether users select the elements that best fit their personality traits or player type, but found indications that this is not the case and that other aspects drive participants in creating game configurations (at least in the conducted study). Towards this thesis, this chapter showed that allowing users to impact the motivational component of systems is worthwhile, as this can lead to beneficial individual effects. The considerations in this chapter add to **RQ2** (see Section 1.4). This chapter focused on individual impacts. In the next chapters, we will consider the live-streaming setting of games and focus on analyzing which impact groups (i.e., the viewers) have on individuals (i.e., the streamer). Additionally, we consider how groups can self-administrate, i.e., how individuals can shape the group outcome.

Chapter 5

Interactivity in Game Live-Streams

In this chapter, we focus on game live-streams by considering the interactive options of viewers and what impact they have on the streams. This consideration adds to **RQ3** (see Section 1.4). To this end, we analyzed the options viewers have in live-streaming channels that are particularly keen to integrate their audience. In addition, we assessed viewers' expectations towards interactivity in general with an online survey. We then present *Helpstone*, a system that realizes a set of novel communication and improved interaction channels. A study with it showed that these have an impact on the streamer and that the added options are perceived positively by viewers. In general, the chapter shows that empowering the individual influence in live-streams is reasonable, as long as the viewers' impact is not unrestricted and can be orchestrated by a streamer.

Section 5.2.1 is based on the publications [165, 171], Section 5.2.2 is based on [168] and Section 5.3 is based on [173].

5.1 Introduction

As illustrated in Section 1.2.2 and Section 2.5, the context of game live-streams is interesting for this thesis, as a group of people (the viewers) watches how an individual (the streamer) plays through games and can potentially exert influence on this person. While an audience always can exert influence implicitly [244] (e.g., when viewer numbers are dropping), we focus on the direct influence options. In Section 2.5.1 we presented current solutions on empowering viewers in live-streams, from a scientific, streamer's and platform point of view. This chapter will add to these ongoing efforts. We have an inclusive view of interactivity: we not only consider options in which viewers impact the streamed content, but also

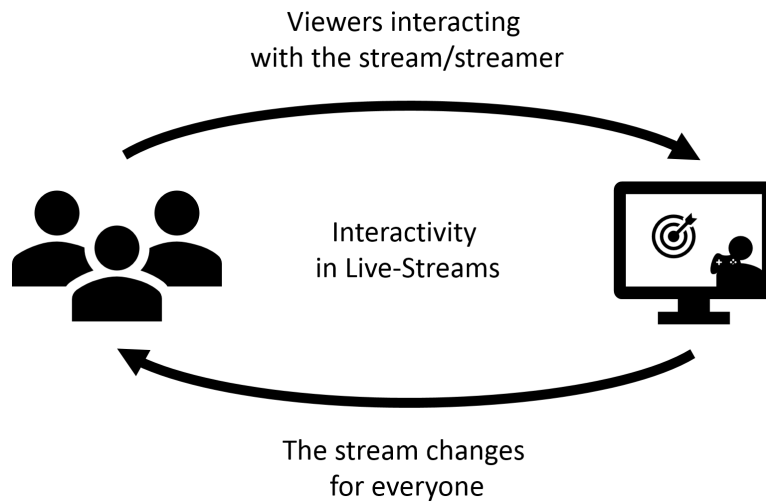


Figure 5.1: Instantiated schematic of reciprocity in interactive live-streams.

options that alter the experience for the viewer him- or herself (e.g., being able to alter which camera perspective is shown). Based on the mentioned issues, such as the chat being a suboptimal communication medium especially in large streams or the fact that consuming live-streams is a shared medium (see Section 2.5.1), this raises the question of how to establish interactivity in live-streams in general. In contrast to the self-sustaining systems (see Chapter 3), where individuals were only loosely coupled (e.g., via leaderboards or implicitly through the improved service outcome), in the live-streaming scenario, the coupling can be considered as high. Everyone sees the same video, and thus influence leading to a change of the streamed content has a direct impact on other viewers as well. Figure 5.1 shows an instantiated schematic of the reciprocity in interactive live-streams. In this chapter, we consider typical live-streaming scenarios, i.e., where at least one streamer is present. In Chapter 6, we complement this view by considering setups without a streamer. In both chapters, our main focus will be on the viewers.

This chapter is structured as follows: first, we consider the interactive aspects of live-streams by analyzing existing channels known for being highly interactive. By doing that we see which influence options the viewers have in live-streaming channels that are keen on integrating the viewers. This is complemented by the presentation of an online study in which we analyzed viewers' expectations and requirements for interactive features in live-streams. Here, we considered not only features available today, but also features that might be options in future live-streaming experiences. We then present *Helpstone*, a system that provides enhanced communication and interaction channels for viewers of live-streams of the popular trading card game *Hearthstone: Heroes of Warcraft*⁴⁰. We present the system and a study we conducted "in the wild" with it.

⁴⁰ Created by Blizzard Entertainment; see <https://goo.gl/Z4Gtk8> (last accessed: 2018-07-07)

5.2 Analysis of Interactivity in Game Live-Streams

While there are different options to support audience interaction (see Section 2.5.1), it is currently unclear how streamers use them in game live-streams and how viewers perceive them on a larger scale. This section addresses this question by discussing interactivity and presenting a case study around a stream format that is known for being highly interactive. We reviewed more than 20 hours of this format for elements that involved the audience and considered how they influenced the content. We will then present a large-scale online survey, in which we assessed viewers' perception of interactive features, and looked deeper into the role of interactivity in live-streams and how much impact viewers want. The studies were done to fulfill these goals:

- Goal_{Interactivity} 1** *Investigation of how streamers integrate their audience today:* By learning how channels integrate their audience, we can derive how much impact individuals have today and are also able to analyze drawbacks of currently used interaction channels.
- Goal_{Interactivity} 2** *Investigation of how viewers perceive interactive features:* The usage of interactive elements does not necessarily mean that they are also perceived as beneficial by viewers. How viewers perceive such features should thus be investigated.

5.2.1 Interactivity Today

As elaborated in Section 2.5.1, games appeared that allow audiences to alter game mechanics while streamers play them. The audiences thus exert direct influence. Also, streams exist in which (originally) analog games are played and even here, channels have started to integrate the audience. An example are games such as *Superfight*⁴¹ that, in their non-live-streaming form, require the players to discuss after having played certain cards to determine who is more convincing. In these games, players have a set of cards and they play a defined number of these per round. In *Superfight*, for example, players have cards representing characters, such as *Abraham Lincoln* and attributes such as *Riding a Segway* or *Long Metal Claws*. After the players have revealed their selected cards, they argue for why their cards are better than the opponent's: in the case of *Superfight*, why their character would win a battle against the other character. After a bit of discussion, the players decide who was most convincing. Such games are also played in live-streaming channels⁴². Here, it is not the other players who need to be convinced, but rather the audience, which can vote via the chat on which player has won (see Figure 5.2). Such games appear suitable for live-streaming, as the audience can be easily integrated and even may improve the game, as their decision is (potentially) less biased than having the players decide.

⁴¹ Created by Jack Dire; see <https://www.superfightgame.com> (last accessed: 2018-07-07)

⁴² For example, see <https://www.twitch.tv/superfight> (last accessed: 2018-07-07)



Figure 5.2: The game *Superfight* on *Twitch*. Left: Voting phase. Right: Result showing how the audience has voted (screenshots taken from <https://goo.gl/eVTKJL> (last accessed: 2018-07-07)).

The core mechanic seen here is based on polls, but further interactive options can easily be imagined: for example, to visualize each player's hand and let the audience vote on which card they want to see played. Another option is to allow viewers to also provide explanations for why a certain card combination should win over the others that are visualized in the stream itself. Finally, the viewer might suggest new card labels that might be integrated into the deck. If all players play with a mobile app (instead of physical cards), these new suggestions could be directly integrated into their games. Thus, they would have to argue using ideas (in the case of *Superfight*, new characters and new attributes) coming from their audience. In general, considering non-digital games appears worthwhile, as here, the underlying system can easily be adapted by streamers themselves in contrast to digital games. Visualizations such as shown in Figure 5.2 are not necessary, but also possible. As these examples show that the range of integrative possibilities appears high, even in such a relatively simple game, we conducted a case study in a stream with a streaming format that uses an analogous game, to learn what is used today, without being restricted by, for example, the programming.

Case Study: *Rocket Beans TV Pen & Paper*

*Rocket Beans TV*⁴³ is a German live-streaming channel broadcasting 24/7. In 2014, they launched a pen & paper role-playing game format, in which the audience is encouraged to participate through various means. Today, the format attracts more than 30,000 viewers during the live sessions and receives many views as video-on-demand. In a pen & paper role-playing game [309] one player (called *the game master*) represents the game world/narrator and can flexibly react to player actions. Players interact within this world in the form of an improvisational theater and can explore the story and the world the game master has prepared. Usually there exist rules to handle character creation and actions inside this world (e.g., fights), and dice are often used to make it more interesting by

⁴³<https://www.rocketbeans.tv> (last accessed: 2018-07-07)



Figure 5.3: Setup of the *B.E.A.R.D.S.* pen & paper session, with the *Twitter Wall* on the left in the background. On the right, a screen shows visualizations fitting the current scenes. Picture taken from episode six, <https://goo.gl/3SFq4e> (last accessed: 2018-07-07).

introducing randomness. In comparison to a video game, the “game engine”, “programming” and “storytelling” are represented by a human. Thus, in contrast to programmed video games, the only limiting factor is the imagination of the people playing this game, making it particularly useful for our analysis. Following this argumentation (i.e., that the content does not restrict the interaction), the aspect that *Rocket Beans TV* encourages the audience to participate and the high number of viewers, we assessed this format as particularly suitable for learning what is done today when streamers want to integrate their audience (adding to our **Goal_{Interactivity} 1**). It is also of interest whether the audience integration options in typical *Twitch* gaming streams mentioned in [108] (see also Section 2.5.1) are also found here. Although pen & paper role-playing games are not “mainstream” for *Twitch*, several channels present such formats. Thus, it can still be considered as a relevant context.

The *Rocket Beans TV* pen & paper session consists of four players and a game master sitting around a table; the scenery is arranged to thematically fit the role-playing setting (see Figure 5.3). Viewers can chat via the live-streaming platforms’ chat (which is not shown to the players but to the game master, and thus only primarily allows information exchange between viewers), but can also post via *Twitter* tweets. Tweets are shown in the studio on the *Twitter Wall / Social Media Wall*, which, in contrast to the chat, is also visible to the players. In addition, it allows pictures and memes to be shared. During a stream, information overlays, music and sound effects that fit the current situation are added to the stream and pre-made clips and pictures are shown to visualize certain aspects, i.e., the game master does not only rely on the imagination of the players and viewers.

Method

We reviewed the first six episode (around 24 hours of video material) of the pen & paper format. This covers all episodes of season one (called *T.E.A.R.S*; a post-apocalyptic zombie setting) and one of season two⁴⁴ (called *B.E.A.R.D.S*; a Vikings setting). At the time of the study, these were all existing episodes. The goal was to analyze elements that involve the audience, following an open coding scheme by using a thematic based analysis [118]. We annotated direct (e.g., the community is encouraged to vote or otherwise directly addressed) and indirect (e.g., the camera shows the *Twitter Wall*, or viewer-created elements are shown) social interactions, with a timestamp and a short description. This transcript was used to derive major themes which were discussed by two researchers.

Results

We counted 209 direct and 293 indirect interactions and clustered these into 21 categories. The number of instances per category varies (with smaller categories such as *viewers providing hints on the game rules* to large categories such as *direct acknowledgments of user contributions*). We related categories and they led to overall themes, which are presented as results next.

Voting: A core element in this format is voting. Until episode five, this was conducted via an external web page. Usually these polls were published just before advertisements were shown. This gave the audience time to vote, without missing any content in the stream (as the transition to the external page was necessary). Due to synchronization problems with the story's progress and the need to display advertisements at specific times, polls were also used during the session. In episode six, the *Twitch* chat was used for voting: the question and answer options (which could be voted on by chat commands) were displayed as a stream overlay. By voting, the audience could decide how scenes should proceed (see Table 5.1). The results were visualized and the most popular answers were used by the game master. In total, 24 polls were conducted and on average 9533 (SD=5338) votes were given. Not considering the first episode, in which the format was tested, and the sixth episode, in which only registered users were able to vote via the chat, the number increased to 13118 (SD=1563). Five times they used polls where viewers could tweet possible answer options. These were screened by people working with *Rocket Beans TV* and they generated a poll based on selected answers. Such polls are more difficult for the game master, as he or she needs to improvise, while for pre-defined answers, scenarios for each outcome could be derived beforehand. As this is still moderated, the audience influence remains orchestrated. In contrast to these live polls, the community was allowed to participate in a poll with pre-defined answers between the two seasons to decide which setting should be played next.

⁴⁴The videos are available on the *Rocket Beans TV YouTube Channel* <https://www.youtube.com/user/rocketbeanstv> (last accessed: 2018-07-07).

Question	Options
What will the group encounter at the bottom of the stairs?	A zombie eating the guard Frank The guard Frank still searching for the key Another prisoner A popcorn machine
What happens in the night?	They will be wakened by scary sounds Their shelter begins to burn Someone calls one of them
What does Steven do?	He eats a leg He and a guest are drinking tea

Table 5.1: Examples of polls used in the *Rocket Beans* TV pen & paper format.

Direct influence on the setting and story: Viewers had the chance to send illustrations and descriptive texts of items the players found within the story. Small cards with representations of the fictive items were given to the player of the character owning the item. The viewer incentive, besides getting directly acknowledged in the stream, was that items were usually available across several episodes and thus were potentially shown multiple times. Before episode six, the audience was asked to send pictures and video material fitting the setting. Selected elements were shown during the episode, and even though they were not relevant for influencing the course of action, they were part of the content shown in the stream. However, the audience also had the chance to influence the story: in episode two, the audience was spontaneously invited to generate a name for a building in the game. This was picked up for the second season, where story elements could be generated collaboratively on an external web page. They could create them freely, or could provide explanations and content for aspects that were already added by the game master. Parts of the content were approved by the game master and then used in the game. The viewers could thereby influence the imaginary world, although the decision on what would be integrated was again not purely audience-driven. Additionally, the audience received tasks to be carried out that were directly interwoven with the story and the world (i.e., if a task was not completed, the situation would worsen for the players). In episode five, the viewers were told to post photos on *Twitter* showing a German landmark with themselves disguised as zombies in front of it. In episode six, the audience represented the inhabitants of a town, and their task was to decide whether they are convinced by a speech given by the players. They were to respond via *Twitter* by sending “thumbs up” or “thumbs down”.

Communication channel for the audience: Players often read tweets on the *Twitter Wall*, especially when they were less involved in a game situation. They even praised the community engagement several times. The *Twitter Wall* is often shown implicitly, when the camera position focuses on specific players, or explicitly, either because the content seemed interesting/fitting for the stage direction or because the players were discussing parts of it. Thus, the wall was directly influencing the content of the stream. Images from the wall were shown and discussed, and comments, either from during or between episodes,

were often read and discussed by the players. The name of the contributor was mentioned or shown in the stream, and the players also acknowledged good contributions directly. The players used suggestions by the audience to alter their behavior in the game, e.g., by asking other story-relevant questions in-game or re-interpreting rules because of a viewer hint. Additionally, help from the audience was also explicitly encouraged by the game master whenever riddles were encountered by the players.

Discussion

This case study revealed different kinds of audience participation options that are used in a stream today that has the goal to incorporate interactive elements. We restricted ourselves to the elements that are shown directly in the stream, not social media sources around the streaming experience that were not directly involved (*Facebook*, *Reddit*, etc.). Nonetheless, all the elements we found are at some point moderated, by either the game master, the players or the team behind the scenes, so they do not offer the audience direct (unfiltered) influence. Through the *Twitter Wall*, viewers have a direct channel in the stream, which is influencing the course of action during the stream. Through the integration of user-generated content, polls and other ways to shape the story, the audience has some kind of shared, but orchestrated, authorship. It can also be seen that single viewers' suggestions are directly incorporated (e.g., user-generated content such as images) and polls suggest what the majority of the viewers want to happen in the story, i.e., they provide influence over the content of the stream in a nearly-real-time fashion. Besides these synchronous actions during the stream, there are also elements that alter asynchronously how the content will change in the future, such as the work on the story and environment, in a collaborative fashion.

We found means for audience influence that were also observed by Hamilton et al. in streams which could be considered more "mainstream" [108]. The difference from our findings is that all options we report were integrated in one stream, while it remained unclear to what extent they were available in the channels considered in [108]. Elements such as the co-story creation option, and directly shaping the experience that unfolds in the stream, were not reported by them. One explanation is that in these channels digital games were played that do not easily allow for such adaptations. The high interest the community shows in the format considered in our study hints that games/streams should offer more of these options. The large overlap of elements here and in the streams considered by Hamilton et al. indicate that there might be only a few common concepts for when the audience should be integrated into the streaming experience today.

Based on the methodology, we do not know yet how these options are perceived by the audience. Maybe the technological considerations are the limiting factor, i.e., concepts for more audience integration are desired but simply cannot be realized with the current setups; or maybe the audience itself is satisfied already, making it unnecessary to develop more. As our case study analyzed only video

content, we were not able to collect viewers' opinions on the different interactive elements. Nonetheless, the study has contributed insights into a particular stream offering integration options for their viewers, and revealed that many options are already at the streamer's disposal. To learn about the viewers' perceptions as a next step, we set up an online study to assess these.

5.2.2 Viewers' Perception of Interactivity

In relation to **Goal_{Interactivity} 2**, we conducted an online study to investigate the following questions:

Q1 Which elements do viewers find interesting while watching game live-streams?

Q2 To what degree do viewers want to be included in game live-streams?

With **Q1**, and by including existing and not-yet-existing elements in the survey, it is possible to reason about what game live-streams should offer for viewers. With it, it is also possible to derive what significance interactive features have. **Q2** provides helpful insights on assumptions that many recent works make only implicitly, e.g., that viewers actually want to be integrated into such live-streams.

Method

We set up an online questionnaire in German and stated that it was only of relevance for people that at least occasionally consume (or have consumed) game live-streams. We defined "game live-streams" as streams in which the streamer actively recognizes his or her audience during the "game performance". We also stressed that "gaming" is not limited to live-streamed digital gaming content (e.g., live-streams showing analog board game play would also be relevant). We collected self-report data on how participants consume game live-streams and integrated questions (e.g., *"In relation to game live-streams, I miss concepts/features on current live-streaming platforms"*) to be answered on a 4-point scale with the labels *disagree*, *somewhat disagree*, *somewhat agree*, *agree*. If they stated that elements were missing, they were presented with a free text question (the different free text questions will be abbreviated with FT subsequently), where they could give details (FT 1). Then we presented ten motivation statements (based on the personas in [41]) for why one consumes game live-streams and they could select multiple ones, followed by the optional free text questions: *"Which concepts/features do you find important on game live-streaming platforms?"* (FT 2), *"Which concepts/features have you already experienced when a streamer wanted to integrate his or her audience in his or her stream?"* (FT 3) and *"Which concepts/features would you appreciate to become better integrated into the stream by a streamer?"* (FT 4). Participants were then confronted with 58 elements related to features, concepts and behaviors in

the live-streaming context. For every element, they needed to state how interesting they would find it in the context of game live-streams (also on a 4-point scale). If they did not know about the element yet, they were asked to state how interesting they would find it in theory. We also integrated a test question where “interesting” needed to be selected, to check that they had actually read properly [197]. Participants could also state in a free text field which elements they found interesting that were not asked about (FT 5). The questionnaire closed with demographic questions and a free text field for any final comments (FT 6).

Establishment of the element set: To establish the set of 58 elements, we did an informal review of today’s major live-streaming platforms and several live-streaming channels, and we conducted a one-hour design workshop with eight consumers of live-streams (aged between 20 and 25 years). In this workshop, we discussed their experiences with audience integration, and which elements they know and which they would find reasonable in the future. Additionally, we consulted the scientific literature in respect to *Social TV*, live-streaming and audience participation, with the goal to identify aspects that are already used today as well as ones that might become relevant. Overall, the outcome (see Tables 5.3–5.6 below) contains features (e.g., availability of a live chat), concepts (e.g., showing what music is played in the channel) and streamers’ behaviors (e.g., acknowledging individual viewers). To assess the expressiveness of the resulting set, we ensured that participants in the questionnaire had multiple opportunities to report on (further) elements in free text questions (FT 1–6). The answers to these and the ranking of our elements provided an overview for **Q1**.

Results

The questionnaire (available in July/August 2017) was promoted via *Reddit* (targeting subreddits for surveys, gaming and live-streaming), *Facebook* (targeting groups for surveys, games and live-streaming of games) and student mailing lists, and by directly contacting streamers with the request to share it with their viewers. Filling out the questionnaire took 21 minutes on average. We filtered responses of participants that completed the questionnaire in under seven minutes, answered the 58 elements with a standard deviation of less than .5 (indicating that they might have only clicked through them) and responses in which the test question was answered incorrectly [197], leading to an answer set of 417 responses (317 male, 93 female, seven no answer; age: <18: 18, 18–24: 180, 25–31: 157, 32–38: 48, 39–45: 9, no answer: 5; 345 reported being German).

Considering the free text questions, FT 1 (which was only conditionally shown to 61 participants (15%)) was answered by 30 (49%), FT 2 by 145 (35%), FT 3 by 179 (43%) and FT 4 by 80 participants (19%). The free text field after seeing our elements (FT 5) was answered by 20 participants (5%). The closing free text field FT 6 was answered by many participants, but only 21 participants (5%) provided a thematically relevant addition to the questionnaire itself. The answers to these free text fields were used to support the found results qualitatively.

Participants: 370 participants (89%) define themselves as a “gamer” as they at least somewhat agreed (i.e., selected either somewhat agree or agree) to this question. Considering how many hours in a week they watch game live-streams, 33 participants (8%) reported 0–1 hours, 108 (26%) 2–3 hours, 151 (36%) 4–9 hours, 81 (19%) 10–18 hours and 44 (11%) reported watching more than 18 hours. Mainly, participants reported watching 1–2 streamers (171, 41%) or 3–4 regularly (160, 38%). Nearly all participants know of *Twitch* (97%) and *YouTube Gaming* / gaming live-streams on *YouTube* (97%). *Mixer* was known by only 107 participants (26%) and *SmashCast* by 58 participants (14%). 221 (53%) report using *Twitch* as their primary platform, 194 (47%) *YouTube Gaming* / gaming live-streams on *YouTube* and one *Mixer*. 294 participants (71%) either had donated, subscribed for payment and/or had already crafted something (a picture, a video, something tangible, etc.) for a streamer. 24 participants (6%) reported being streamers themselves and 96 streamers were mentioned as answers to the question “What is your favorite streamer?”, with *Rocket Beans TV*, *Gronkh* and *Bonjwa* being named most often.

Expressiveness of the element set: We conducted an open coding content analysis [118] of the free text answers (FT 1–6), to see which elements were mentioned and to assess the expressiveness of our element set. While 40 elements that were formulated there were also covered by our set of 58 elements, 15 elements were mentioned that we did not have (see Table 5.2). Overall, we reason that the most often mentioned ones should also be included in future iterations of similar questionnaires. In addition, five aspects that were mentioned multiple times were ones that we see as necessary prerequisites in streams, which is why we did not integrate these in our set before (having a good streaming quality (15×), well-designed overlays (1×)); that we integrated with other questions (the streamer needs to have a likable personality (33×), third-party tool functionality should be available directly in the platforms (4×)); or that concerned “meta” elements of the platforms (having a good usability, even for gaming consoles; ensuring privacy; having no advertisements and no rules for the streamers (14×)).

We let participants rate our set of existing/not-yet-existing elements to learn how these elements are ranked. For presentation reasons in this section we use different tables: we clustered the elements into general aspects for the live-streaming experience and live-streaming platforms (see Table 5.3), aspects that allow the audience to influence the stream (see Table 5.4), elements that relate directly to the streamer’s behavior (see Table 5.5) and aspects that relate to visual/auditive elements in the stream (see Table 5.6). Every table provides the overall rank of the element (based on their agreement rating, i.e., how many participants rated the feature as “somewhat interesting” or “interesting”), an indication of whether the feature was mentioned by at least one participant in the preceding free text questions and the agreement rating itself.

By considering the element ranking throughout the different tables, the answers to the other questions and to the free text questions, we derive the following main results, which are elaborated on in the next sections:

ID	Elements mentioned in the free text fields	#
A01	A replay functionality to re-watch specific aspects, potentially also directly in the stream, to be triggered by viewers	12
A02	The ability to like/follow/subscribe streamers	8
A03	Direct availability of videos after live-streaming (VOD)	8
A04	An easier way for viewers to play in community games (an automatic selection of viewers of the channel and direct adding to the game)	4
A05	Streams should have (or not have) a regular schedule	3
A06	The streamer is visible during streams	3
A07	Viewers can formulate missions for the streamer that he or she needs to fulfill in the game/in the stream	2
A08	Having an option to watch streams in VR	2
A09	The streamer comments on games which are played by his or her audience	2
A10	Seamless integration of live-stream and VOD, i.e., continue the VOD at the point where I have left the live-stream	1
A11	Questions already answered via chat should be automatically posted when the question is asked again	1
A12	The streamer requests his or her audience to visit another channel ("raids")	1
A13	A viewer should be able to customize the overlays shown by a streamer for him- or herself (i.e., suppress donation trackers).	1
A14	A viewer should be able to take complete control over the game the streamer is playing for a short time	1
A15	Betting with real money	1

Table 5.2: Elements we found in the free text fields that are not integrated in our set. # denotes the number of participants that mentioned the element.

- R1** Overall, users are satisfied with the elements available in game live-streams today, yet some could improve
- R2** Interactive and interactivity-related elements received high ratings
- R3** Audience integration is relevant, even for passive viewers
- R4** Trolling, bad past experiences and context factors are challenges for audience integration
- R5** Different viewer motivations exist and have a moderating effect on the element perception
- R6** The impact of the audience should not interfere with the streamer's performance unconditionally
- R7** Established streaming behaviors should be revisited

ID/ Rank	Free texts	Elements relating to the streaming experience and general features in streams	Agree- ment
01	✓	Anti-trolling mechanics	91.8%
03	✓	No delay (lag) between streamer and viewer	87.3%
05	✓	A chat bot to query channel-related information (e.g., current uptime)	85.9%
10	✓	The channel description	75.8%
12	✓	Multiple camera perspectives; every viewer can change the perspective for him- or herself	73.9%
14	✓	The live chat	70.3%
17	✓	Availability of channel-specific emoticons	61.6%
18	✓	A chat bot that writes meta-information on the current game into the chat (e.g., win/loss ratio of the streamer in this game)	60.7%
20	✓	Being able to upvote individual chat messages, which then remain visible for longer	58.8%
21	✓	Availability of standard emoticons	53.5%
22	-	To have more information on the current game as viewer than the streamer in a stream (e.g., seeing enemy positions)	52.5%
23	✓	360 degree video stream; every viewer can manipulate the perspective for him- or herself	51.3%
24	-	Automatic extraction of chat topics that are shown to streamer and viewers together with the latest messages on that topic	50.6%
26	-	Viewers that watch the streamer more often have additional features	46.3%
29	-	Channel-specific achievements can be unlocked (e.g., after taking part in many polls) that are visible to all other viewers	43.4%
31	✓	A betting system and a virtual currency to bet on the outcome of games in a channel	41.7%
35	✓	Access to additional features for subscribers of the channel	40.5%
36	✓	Gamification elements for viewers (e.g., a virtual currency that increases the longer a viewer watches a stream)	40.5%
38	-	To provide comments in the channel, even if the stream is offline	39.1%
39	-	Enable subtitles in your language	37.6%
41	✓	Mini games that can be played in the live chat in parallel to the stream	36.2%
44	-	Multiple camera perspectives; which perspective is shown to all viewers is decided by an ongoing poll	35.3%
51	-	The option to chat privately with other viewers ("whispering")	28.1%
53	✓	Availability of chat rooms	24.5%
54	✓	An automatic classification of viewers, and viewers in the same class will be put in the same chat room	23%
56	-	360 degree video stream; which perspective is shown to all viewers is decided by an ongoing poll	21.8%

Table 5.3: Elements related to the streaming experience or general features. The first column indicates the overall rank (based on the agreement score) and serves as ID. The column "Free texts" indicates that at least one participant suggested the feature in one of the four preceding free text questions. The column "Agreement" shows the percentage of participants (bold if larger than 50%) that rate the feature as "somewhat interesting" or "interesting".

ID/ Rank	Free texts	Elements that allow viewers to alter how the stream proceeds/to interact directly with the streamer	Agreement
02	✓	Polls during a stream that are set up by the streamer	89.9%
08	✓	Polls between streams that are set up by the streamer	81.3%
09	✓	Being integrated in the game the streamer plays, e.g., in a quiz game, to play along by also answering in the chat	79.4%
16	✓	Polls during a stream where viewers can add answer options	62.1%
19	-	Polls between streams where viewers can add answer options	59%
25	✓	Viewers can change game elements of the game the streamer is playing (e.g., changing the kind of monsters)	50.1%
27	✓	Viewers can change the difficulty of the game the streamer is playing (e.g., taking away the current weapon)	45.3%
30	✓	Viewers can send virtual items or provide other assistance for the game the streamer is playing (e.g., providing more ammunition for the current weapon in an ego-shooter)	42.4%
32	✓	Viewers are able to change the background music in the stream (e.g., with a poll)	41.2%
37	✓	Viewers can interact with the streamer directly, e.g., with buttons below the video stream	40%
40	✓	Viewers can directly interact with the video stream to provide hints to the streamer (e.g., by drawing lines onto the streaming window). An aggregation system aggregates the same hints	37.2%
49	-	The option to set up and start polls as a viewer	30.5%
50	-	Viewers can manipulate the streamer's gaming setup (e.g., swap keybindings) for a short time	29%
55	✓	The viewers can decide how individual votes will be combined (not only majority votes)	22.3%

Table 5.4: Elements that allow viewers to alter how the stream proceeds/to interact directly with the streamer.

ID/ Rank	Free texts	Elements related to the streamer's behavior	Agreement
04	✓	The streamer reacts to chat messages in the stream	86.8%
07	✓	Viewer games (the streamer plays with or against his community)	81.8%
13	✓	The streamer shows user-generated content (e.g., pictures) in the stream	73.6%
15	✓	The streamer plays viewer-submitted modifications (e.g., a mod for <i>GTA V</i>) or specific content (e.g., a map for <i>Minecraft</i>)	64.5%
28	✓	The streamer thanks/acknowledges viewers directly in the stream (e.g., after a donation)	44.1%
34	✓	The streamer shows selected comments from social media platforms directly in his or her stream (e.g., showing <i>Facebook</i> posts)	40.5%
42	✓	The streamer does raffles or distributes giveaways	36%
43	✓	The streamer adds viewers via <i>TeamSpeak/Discord/Skype</i> to the stream live	35.5%

Table 5.5: Elements related to the streamer's behavior.

ID/ Rank	Free texts	Elements related to the screen/audio composition of the stream	Agree- ment
06	-	Usage of game-specific overlays that convey additional information (e.g., cards trackers)	82%
11	-	An overlay showing which music is currently playing in the stream	74.6%
33	-	Viewers can submit user-generated content (e.g., pictures) that are automatically shown in a dedicated area in the stream	40.8%
45	✓	Notifications shown in the stream after a viewer takes specific actions (e.g., donating or subscribing)	35%
46	-	Bio signals of the streamer are permanently shown in the stream	34.8%
47	✓	Permanent integration of social media platforms in the stream, e.g., tweets to a <i>Twitter</i> account are always shown	33.6%
48	✓	An always-visible donation tracker in the stream	30.9%
52	✓	Permanently seeing the live chat in the stream	27.3%
57	-	Viewers can record voice messages and submit to the streamer so that they will automatically be played	11.8%
58	-	Mood emoticons that are directly shown in the live-stream	10.6%

Table 5.6: Elements related to the screen/audio composition of the stream.

Overall, users are satisfied with the elements available in game live-streams today, yet some could improve (R1): Before participants were presented with our element set, we asked participants if any elements are missing from live-streaming platforms. 233 participants somewhat disagreed and 121 disagreed (summing up to 85%). This shows that the elements used today seem sufficient for the majority of our participants. Although other studies have revealed shortcomings (e.g., [277]), overall the participants seem to be content with what is offered today in this context when asked from a general viewpoint.

Considering the ratings for elements in Table 5.3, we see potential for improvements: adjusting the camera options (*360 degree video stream* (rank 23) or *having multiple perspectives* (12)) is something viewers want, but is only easily possible on *YouTube* (currently). *No delay between streamer and viewer* (03) is only possible on *Mixer* (see the lag issue explained in Section 2.5.1). The option *to upvote individual chat messages that remain visible for longer* (20) and the *automatic extraction of chat topics* (24) are highly-rated features which are, to our knowledge, not yet available on live-streaming platforms, and thus might be valuable additions. Participants are not satisfied with the current communication options, which they expressed in the free texts⁴⁵:

“The chat is currently very restricted... Many streamers use chat bots for the IRC but this feels like the Internet stone age.”

“A better audience integration would be achieved if there were a better overview in the chat.”

⁴⁵ Participants’ free text statements were translated from German to English.

“Sometimes a slow mode for the chat, so as not to miss important answers, would be good if many people are in the chat”

By further considering Table 5.3 (the other tables will be considered in the next results) apparently unnecessary general features can also be seen: *comments when the channel is offline* (38), the option to *chat with other participants directly* (“whispering”) (51) and the availability of *chat rooms* (53) are elements that are available today, but are rated as uninteresting, showing that these features do not add much to the experience for users. *Having subtitles in the viewer’s language* (39) seems not to be interesting for participants (most likely as participants, from an entertainment perspective, would select streamers they can understand), and the *automatic classification of viewers and moving classes of viewers to the same chat rooms* (54), as a not-yet-existing feature, is also rated as uninteresting, most likely as the concept of chat rooms is not liked. Perceptions of gamification options (29, 31, 36, 41), which are currently being offered in part by the platform *Mixer*, are also mixed (as every element scored below 50%). This is interesting, as 89% of our sample reported to be gaming-affine. We hypothesize that in streams such elements are not as important as the stream itself, which is already interesting enough. Thus, motivational elements on top of it seem unnecessary.

Interactive and interactivity-related elements received high ratings (R2): Considering the complete ranking of the elements, we see that 15 of the top 20 elements (IDs/ranks 01 to 20) are directly related to interactivity:

- *Anti-trolling mechanics* (01) received the highest rating. Trolling [39] impacts the streaming experience [270] and is an issue for interactivity (see **R4**).
- *Polls set up by the streamer during* (02) and *between streams* (08) are interesting for many participants. 146 participants also mentioned polls in the free texts, showing that this is a well-established element (fitting to our previous study). *Polls during a stream* (16) and *between streams* (19) where *viewers can add answer options* are also in the top 20.
- *Having no delay between streamer and viewer* (03) is important for interactive concepts. The high ranking is also an indication that participants (as only a minority know of *Mixer*, which has overcome this issue already) did not only rate elements high when they already knew about them.
- *Communication between streamer and the chat* (04) was also mentioned 61 times in the free texts and can be considered as a basic interactive concept.
- Having the option to *play with the streamer (or against him or her)* (07) or being *integrated in the game the streamer plays* (09) was also mentioned several times in the free texts (46 and 49 times, respectively).
- *Multiple camera perspectives* (12) allow viewers to adapt the stream view to their interests. Potentially, with such a feature, interactivity can be further enhanced, for example when viewers can focus on specific parts of a stream on which they can exert influence.

- Streamers that *show user-generated content in the stream* (13) or *playing viewer-submitted modifications or specific content* (15) are showing integrative behaviors which were also rated highly.
- *The live chat* (14) is, as described, the primary interaction channel today.
- *Availability of channel-specific emoticons* (17) can also be considered an interactive element as work such as [221] showed the relevancy of emoticons, especially in large channels, as successful means for communication.
- *Being able to upvote individual chat messages, that then remain visible for longer* (20) makes good contributions more distinct for other viewers and the streamer, and thus mitigates effects of information overload and helps to improve the interaction between both parties.

The amount of interactive and interactivity-related features in the top 20 is an indication that interactivity is important for viewers in live-streaming (see also Section 2.5). As some features might be hard to realize directly on a live-streaming platform when not implemented by the vendors themselves, we asked whether participants would be open to move to an external web page where live-stream and chat were integrated. We received mixed answers, as only 225 (54%) agreed to this at least somewhat. The same is also true for whether they would install a browser plugin (211, 51%). Taken together, it appears that the live-streaming platforms need to offer novel aspects directly to maximize their value for viewers.

Audience integration is relevant, even for passive viewers (R3): We asked the 24 streamers that took part in our questionnaire how important the audience integration is for them while streaming. All but one reported that it is at least somewhat important. From the audience perspective, we learned that 294 (71%) agreed at least somewhat to the question whether they like game live-streams where they are integrated as a viewer (for example with polls). Through the answers to the free-text fields, we learned that “being integrated” starts even with “simple” interactions between streamer and his or her audience, exemplified by the following statement of a participant:

“Interactivity is important. I like it when a streamer talks with me and his audience. I really appreciate it because it feels like I am sitting on a couch with friends. I also like it when the audience is able to decide whether a game should continue streaming or not.”

By considering the interactive elements the majority of our participants assessed as interesting (see **R2**), we see that many of these are rated as interesting even by more than 71% of the viewers. This hints that participants might have a different understanding of what it means “to be integrated” in game live-streams (as some disagreed to liking being integrated, but rated integration features as interesting for them). Other participants provide statements that reveal that they do not want to engage themselves as viewers in streams, but still appreciate if the streamer integrates his or her audience:

"I don't care whether I am being integrated, but I like it if a streamer does this, as it provides variety."

"For me, the most important feature is the integration of the community, even though I would not participate myself."

"Although I don't use the chat much, I think the chat is the most important component for live-streaming, as I appreciate reading what is written there."

Our data indicates that this appears to be true for many participants: 335 (80%) agreed at least somewhat to the statement that they are passive viewers and would, for example, not use the chat actively. Such a high number of passive viewers was also reported by Gandolfi [83]; thus it is not only our sample, but seems to be a more general case. Additionally, 357 participants (86%) reported that they are not really interested in communicating/interacting with other viewers in game live-streams and 291 (70%) even stated that they are not interested in communicating/interacting with the streamer. This is a surprising result in respect to the related work where the social aspects and community shaping was found to be an important topic in live-streams (see Section 2.5). We compared participants who claimed to be passive and do not want to interact with other viewers/the streamer (265 participants (64%) who will be described as "passive viewers" subsequently) to participants that provided at least one positive answer to one of these statements (152 participants (36%), "active viewers").

Of the 265 passive viewers, 161 (61%) agreed at least somewhat to the statement that they want to be integrated ($M=2.6$, $SD=.9$, $Mdn=3$); of the 152 active users, 133 (88%, $M=3.3$, $SD=.7$, $Mdn=3$). A t-test revealed this to be a significant difference with a medium effect size ($t(371.9)=8.4$, $p<.001$, $r=.4$). Nonetheless, as even the majority of the "passive viewers" also like it when viewers are integrated, it shows that even though they might not want to participate in such interactive options, they see a certain appeal to them. Taken together, we conclude that the integration of viewers is a relevant topic for live-streams today, for active and passive viewers alike, and that "being integrated" already starts when streamers acknowledge their viewers. We additionally considered whether there is a difference between passive and active viewers on the question of whether any features are missing on live-streaming platforms ($M=1.8$, $SD=.7$, $Mdn=2$ vs. $M=2$, $SD=.8$, $Mdn=2$). A t-test shows this to be a significant difference, although with only a small effect size ($t(425)=-2.6$, $p<.01$, $r=.13$).

Trolling, bad past experiences and context factors are challenges for audience integration (R4): Still, 29% of our sample somewhat disagreed or disagreed to the statement that they like game live-streams where they are integrated as a viewer. One reason we identified in the free text fields, and the other answers that can be related to this as an explanation, is that several statements addressed trolling behavior [39] of other viewers as a problem:

"I find features that manipulate the stream uninteresting. There are too many trolls and spammers."

"In general, I am not a fan of things that affect the streamer. Often there are trolls..."

"I'm of the opinion that too many features lead to issues: backseat gaming or trolls. Additionally, there could be delays and the flow of the game could suffer."

That trolling is a severe issue for game live-streams is also supported by the fact that 293 participants (70%) agreed at least somewhat to the statement that trolls are annoying for them during a game live-stream and by the fact that *anti-trolling mechanisms* (01) were ranked highest. Negative experiences with integrative options were also mentioned in the free text answers:

"I think streamers should do what they want. Always these polls... 40% are against it, 60% want it and in the end many are angry because they have not received what they want."

"I don't want to be integrated – in the end, it is always bad."

An important aspect that was also revealed was that interactive options appear to be context-dependent:

"Interactive games like "Quiplash" or "Choice Chamber" are great but not permanent."

"Influencing the game of the streamer is only interesting if the game itself has mechanics for this (e.g., "Choice Chamber" or "Party Hard"). Games such as "Call of Duty" are unsuitable for such concepts."

"Please do not overuse interactivity. Too much of it is not good – when used discreetly it helps to increase the entertainment value and the stream overall, but if it is omnipresent, I lose interest. I watch streams in parallel to playing games; thus I cannot click every five minutes on something on my second screen."

"Many of the proposed concepts are funny shenanigans for a short time. From a long-term perspective these are probably nerve-racking. Good for short events or for streamers with a younger target group."

As our questionnaire already was extensive, we had not integrated additional element evaluations for specific scenarios, and took a general viewpoint instead. These statements reveal that a specific perspective might change the perceptions of certain elements. From the qualitative answers, we see that time-wise usage and context factors, e.g., which games are played, are of relevance for interactivity.

Different viewer motivations exist and have a moderating effect on the element perception (R5): Only two participants stated that none of the ten presented statements on the motivations for why they watch live-streams (loosely based on

No.	Statement	Affected element IDs	No. of times selected
M1	It might happen that I watch game live-streams of games that I have never played before and I do not have any clue about. (<i>The Uninformed Bystander</i>)	13, 38, -57	394 (94%)
M2	It might happen that I watch game live-streams of games that I used to play, but I do not play them currently or do not want to play them anymore. (<i>The Uninvested Bystander</i>)	35	381 (91%)
M3	I watch game live-streams to close knowledge gaps about the game and to learn, for example, new strategies. (<i>The Curious</i>)	06, 07, 12, 21, 26, 35	248 (59%)
M4	After watching game live-streams, I am often motivated to play the game and/or try out strategies I have seen in the stream. (<i>The Inspired</i>)	07, 09, 11, 20, 22, 24 to 27, 29, 30, 32, 35, 38, 40, 41, 53	305 (73%)
M5	I watch game live-streams to learn strategies to improve my skill in this game. (<i>The Pupil</i>)	03 to 07, 14, 28, 31, 35, 36, 43, 45, 53, 57	194 (47%)
M6	I watch game live-streams as a substitute for not being able to play the game, for example because I do not own the game or my hardware is not sufficient for it. (<i>The Unsatisfied</i>)	02, 07, 08, 13, 15, 34	272 (65%)
M7	I watch game live-streams to be entertained (similar to television) without putting much effort into it. (<i>The Entertained</i>)	20, 24, -43, -57	378 (91%)
M8	I watch game live-streams because I want to assist the streamer during the stream (e.g., giving hints, being a moderator, ...). (<i>The Assistant</i>)	02 to 04, 07 to 09, 14, 17, 21, 24 to 37, 39 to 45, 49 to 52, 55 to 57	65 (16%)
M9	I watch game live-streams to comment on what I see and to share my knowledge. (<i>The Commentator</i>)	02 to 04, 07 to 09, 13, 14, 16, 17, 19, 21, 22, 24 to 33, 35, 36 to 52, 55 to 57	38 (9%)
M10	I watch game live-streams because I know that many other viewers are there and I can interact with them. (<i>The Crowd</i>)	02 to 05, 07 to 09, 13 to 21, 24 to 37, 40 to 45, 47 to 53, 55 to 58	62 (15%)

Table 5.7: Motivation statements are loosely based on the viewer types (given in parentheses) presented in [41]. Multiple selection was possible. Column 3 enumerates the IDs of elements that had a significantly ($p < .05$) better (or worse, denoted with a -) score when the statement was selected.

the viewer personas in [41]) fit (see Table 5.7). On average, 5.6 statements were selected. It appears that there are many driving factors for viewers. Considering the statements further, many participants are keen to learn new strategies or want to improve their own skill (M3, M5), indicating that streamers who also explain why they do certain moves in the game could spark more interest. Second, even more participants are motivated to play the game after they have watched it (M4). This is relevant for many game developers and vendors [277]. In this sense, M1 and M6 are also relevant, as the majority of our participants claimed that they

watch game live-streams as a substitute for not being able to play the game, and that they also watch games they do not already know. Third, not surprisingly, the main motivation for people to watch live-streams is to be entertained without the need to put any effort in it (M7). This is in line with related work such as [95]. Finally, M8, M9 and M10 are statements fitting viewer types that would benefit the most from a better viewer-streamer integration. In these cases, only a small portion of our participants characterize themselves as motivated by this (fitting the passive viewers in **R3**).

We analyzed relations between the motivation statements and our elements (comparing participants who selected them with those who did not). We were able to find several significant differences (all at least at the $p < .05$ level). The third column in Table 5.7 contains the affected IDs. By inspecting the affected elements, some expected differences were found, e.g., that participants that selected *the Pupil* (M5) statement provided a higher score for *having no lag* (03); that the streamer *reacts to chat messages* (04), *uses game-specific overlays* (06) and *does viewer games* (07); the *live chat* (14) and *including viewers in the stream live* (43). These elements seem in line with the statement as these help the goal of *the Pupil* to learn new strategies to improve their own skill (e.g., when the streamer reacts to questions either in the chat or live). Nonetheless, other relationships are not so obvious. As our statements and the viewer types presented in [41] have no validated connection, this mainly serves as an illustrative example which we will not elaborate upon further at this point. As the focus of this study was on general aspects and not to develop measurements to classify viewers, it seems acceptable to learn that there indeed seem to be different viewer groups that have an effect on the perception of features, concepts and streamers' behaviors.

The impact of the audience should not interfere with the streamer's performance unconditionally (R6): The role of the streamer was highlighted:

"I like good and authentic entertainment (the streamer needs to have fun playing the game)."

"I watch streams because of the streamer and not because of the other viewers."

Overall, 33 participants (8%) reported in free text answers that the personality and enthusiasm of the streamer is very important for them in game live-streams (also reported in [232]). This is also supported by the fact that 308 participants (74%) agreed at least somewhat to the statement that streamers are more important than the games they are playing. In combination with the motivational statement selected by most participants (M7, "entertainment without effort"), that an interaction/communication with the other viewers/the streamer is not so relevant for the majority of the participants (see **R3**), and that the audience wants to identify with the streamer (which was reported in [121]), we conclude that the streamer as person/performer is most important. Thus, extensive audience integration might impact the streamer's performance too much, which was also further expressed in free text statements:

“I don’t want to be integrated, I want to consume and get up when I want while the stream is still working. I want to watch someone play and I do not want to play for him; otherwise, I could also play on my own.”

“When watching, I want to see the streamer playing and how he interacts with the chat verbally. I don’t want to see the chat manipulating the game or the stream.”

Considering the elements for audience integration that were highly rated (see **R2**) vs. those which were not, we see that the former are “moderated” by the streamer. Here, the streamers already know beforehand what might happen and how big the impact will be: *polls set up by the streamer* (02, 08), *viewer games* (07, 09), *reacting to the chat* (04) and *showing/using user-generated content* (13, 15) are moderated by the streamers. The *polls in which viewers can add answer options* (16, 19) are rated worse, but are still rated as interesting by the majority of the participants. We hypothesize that there is still some form of moderation in place: after seeing the poll result, it is still up to the streamers to react to the poll. If one of the newly added answers is not a good fit for them, they can discuss this and select another option or interpret the result to be a better fit for him- or herself.

In contrast, other elements in Table 5.4 that aim at an audience integration which would impact the streamers in an unmoderated fashion (i.e., as the effects occur automatically, a streamer cannot prevent what happens) were rated worse by the participants: *changing game elements* (25), *making the game more difficult* (27), *providing assistance for games* (30, 40) and *manipulating the streaming setup of the streamer* (50) alter considerably how the game in the stream proceeds. Thus, they have an impact on the streamers and their performance. *Changing the background music* (32) and the option that *viewers can set up polls* (49) are also rated worse. Potential reasons for why a lack of moderation here is problematic is that the games then become more chaotic/easier/harder, which could affect the entertainment value for the viewers who are interested in the skills the streamer shows in the (unaltered) game. Also, the danger of trolls increases if such interactions become possible (see **R4**).

Established streaming behaviors should be revisited (R7): We found aspects that are established in streams today but were not rated high in general. Table 5.5 shows behaviors that are done by streamers, but only half of these were rated as interesting by the majority. *Acknowledging viewers (e.g., after a donation)* (28), *showing comments from social media platforms* (34), *doing raffles or distributing giveaways* (42) and *adding viewers via TeamSpeak/Discord/Skype* (43) are not. It also became obvious that many elements of Table 5.6 are rated as somewhat uninteresting or uninteresting and also belong to the worst-rated features in our set: while seeing *an overlay showing which music is playing* (11) in the stream is rated high, elements that are often used in streams today, for example *notifications after events (such as a notification that there is a new subscriber)* (45), *a donation tracker* (48) or *replicating the live chat* (52) are rated worse. We see the potential distracting character of these elements, especially in larger channels.

The same might serve as an explanation for why the other (not yet established) features in this category were not rated well: *an area in which submitted user-generated content is automatically shown* (33), *the permanent integration of social media channels* (47), *the option to submit voice messages that are automatically played* (57) and *emoticons that are “flying” through the stream* (58). Additionally, the chance for trolling behavior is higher. Interestingly, seeing *bio-signals of the streamer* (46), which might spark interesting discussions among the viewers, was rated low as well. Finally, elements that restrict the individual options of viewers are also not perceived well, i.e., every viewer wants to have the same possibilities: *giving more features to viewers who often watch the stream* (26) or *subscribers* (35), not providing viewers with the option to adjust the camera perspective for themselves (12/23 vs. 44/56), *moving viewers automatically into chat rooms with viewers that are similar* (54) and providing viewers with the option *to change how individual votes in polls will be combined* (55) (potentially empowering subsets of viewers) are all rated as interesting by only a minority. Considering the first two mentioned features (ID 26 and 35) these are already done today: the platform *Mixer* provides a currency (that can be spent on actions in the streams) the longer a viewer watches streams and subscriber-only features (such as special emoticons) are used often today. These results show that established behaviors and elements used in streams today should be revisited.

Discussion

Our questionnaire provided insights into current live-streaming consumption behaviors. We learned that the elements that are offered and used today by streamers are sufficient for viewers, as the majority indicated nothing is missing. The resulting element ranking shows that the top elements are already (mostly) available today and that they appeal to a broad range of viewers. Nonetheless, some of them are either not available on every major platform (e.g., no delay), need to be added via third parties (e.g., chat bots) or have potential to be improved (e.g., the communication options) (**R1**). Our element presentation and the elements' ranking can serve as a guide. This aspect also shows that participants did not give high ratings only to elements they are already familiar with. We also found that elements that restrict individual options are problematic, relating directly to the scope of this thesis. Additionally, some features and behaviors in streams that are commonly used today are not attractive for the majority of our sample, and their use should be revisited (**R7**). These findings add to **Q1**.

We also found that amongst the top-rated features, many are interactive, showing that viewers of game live-streams indeed consider interactivity as important (**R2**). Nonetheless, our sample characterized itself as passive, which is not surprising considering the literature on “lurkers” in online communities [127, 240], in which up to 90% of the users are only consuming (in our case, the stream) and the remaining 10% add to the content (in our case, to participate in the chat, for example). We could show that passive viewers still appreciate when interaction

happens, and many integrative options were assessed as interesting by active and passive viewers alike (R3). We also had evidence showing the importance of the streamer and that our sample found “integration features” interesting as long as the streamer is able to moderate what happens (R6). These findings add to Q2.

We also learned that interactive elements should not be overused in streams, and thus specific features might be perceived differently depending on context. From our results, we derive that time and the streamed game/game genre are relevant factors (R4). For this study we decided to take a general viewpoint, i.e., we have not contextualized the elements for specific streaming situations, which had already resulted in an extensive questionnaire. Future work should use specific scenarios (e.g., cooperative or competitive games), analyze to what degree these impact the features and derive guidelines from this. This work can serve as a starting point to identify relevant features that should be included in such a consideration. We also found evidence that besides contextual factors, individual factors are also relevant, as the viewer motivations have a moderating effect on the perception of features (R5). These findings add to both Q1 and Q2.

Regarding limitations, it needs to be kept in mind that we assessed “a priori” expectations, i.e., we have not presented specific feature implementations and let participants assess them. They only received a textual representation and potentially needed to imagine the feature. While this helps to avoid introducing a bias with a fixed scenario/implementation [227] (i.e., participants are able to assess the concept, not the realization), it has the potential drawback that features might be hard to judge by participants when they could not imagine how such elements would unfold in a stream. Thus, elements might be perceived differently in streams than anticipated in our ranking. The assessment is still of value to learn about the viewers’ general perspectives, which can be used to contrast with the context-specific considerations observed in the future. A further aspect to consider is that we asked viewers how *interesting* they find the elements. We assessed this adjective as more inclusive than, for example, using *enjoyable* (as it might be unlikely, for example, that people would say they *enjoy* anti-trolling mechanisms) and as a prerequisite for *appreciating something*. Nonetheless, the wording needs to be kept in mind when working with our results. Another limitation is that we cannot say what impact it has when a streamer uses one or more elements in his or her stream that received only a low scoring, i.e., we cannot state whether it leaves a viewer who marked it as uninteresting simply unmotivated to use it, provokes negative reactions or prompts them to leave the stream. This is also an interesting aspect for future work. Another potential limitation is bias for certain streamers. Many participants mentioned the same streamers as favorites or as being most integrative for the audience. This is a consequence of the way we promoted the questionnaire and might have affected the responses. Finally, as we restricted the sample to German-speaking participants, it is unclear how the results map to other nationalities. Considering just one culture area in live-streaming research is not unusual (e.g., [182]) and considering cultural differences (e.g., [82]) it seemed reasonable to focus first.

5.2.3 Contribution to the Thesis' Questions

The *Rocket Beans TV* case study and the online study revealed important aspects regarding the overall thesis questions (**RQ3**; see Section 1.4). We found that streamers that are keen on integrating viewers use various means to integrate the audience's opinion using existing communication channels, third-party tools and concepts to involve the audience (**Goal_{Interactivity} 1**). Thereby viewers have no unrestricted impact, as at some point their influence is moderated by the streamers. As found in the previously presented online study, this is something that is appreciated by the viewers: the role of the streamer is judged as important and viewers do not want to interfere with the "performance" of the streamer him- or herself unconditionally, i.e., the streamer should be able to orchestrate his or her audience. Besides the streamer's personality, clearly, the danger of trolling is considered as high when more interactive options are available.

Through the online study, we also found that interactivity in live-streams is important to viewers. This already starts when streamers acknowledge their viewers, and more sophisticated features and concepts are perceived as valuable by many viewers (**Goal_{Interactivity} 2**). The entertainment aspect of consuming such game live-streaming performances was shown to be highly relevant for viewers. At the same time, viewers want to have at least the option to have more impact than simply watching how the experience unfolds. We also found individual differences in the perception of specific features, i.e., not everyone wants to use the interactive options to the same degree, but even participants that reported being passive in streams (i.e., those that reported not to want to contribute) also appreciated interactivity. For them, it seems that the live-streaming experience might be enhanced further if other viewers have a (moderated) impact.

Overall, providing viewers in game live-streams with more influence options is reasonable. However, the online study revealed that systems should always consider users that do not want to use their choices. Additionally, through the tight coupling of viewers (and streamers), it became obvious that an unrestricted influence is not valuable. With *Helpstone*, we will present a system in which we enhance the interactive and communication channels in a live-streaming setting, while still allowing the streamer to orchestrate the viewer contributions. With this, we are able to evaluate a set of interactive features when they are realized and used in a live-streaming context.

5.3 Enhancing Interactivity in Game Live-Streams

We aimed at creating a setup in which viewers receive enhanced interaction channels to potentially influence the streamer. Based on the previous finding that streamers should be able to orchestrate this influence, we also did not allow the participants an unrestricted impact. Overall, we had the following goals:

Goal_{HS} 1 *Creation of a system offering improved audience integration options:* As shown in the previous studies, we learned that the existing communication channels can improve and that viewers appreciate more interaction options. Thus, with this goal, we want to offer a system that enhances these channels and allows viewers, if they want, to have an impact on the stream.

Goal_{HS} 2 *Evaluation of improved audience integration options in a live-stream:* By using the system resulting from the previous goal within a user study, we can investigate how it is perceived by a streamer and viewers. We can also investigate what impact the audience can have on the streamer (and thus, on the stream's course).

5.3.1 Concept and System Design of *Helpstone*

To account for **Goal_{HS} 1**, we decided to focus on one game, in which a set of further audience integrative means is implemented and offered. To maximize the validity of later studies with the system, the system should work on current live-streaming platforms. This, on the other hand, also meant that we need to cope with issues such as the lag on this platform, which also guided our game selection. In this sense, for this first investigation, we decided not to use a fast-paced game. At the time of the system's creation (2016), the *Twitch* extensions concept was not yet published and *Mixer* was not popular (see Section 2.5.1).

Following game popularity statistics⁴⁶ at the time of the system's creation, the most popular streamed game which was not fast-paced was (and still is today⁴⁷) *Hearthstone: Heroes of Warcraft*. It was released in 2014 and is a round-based video card game. Players build their deck from a number of cards and play against other players. Every turn, a player has a limited amount of resources (more resources become available every round). A player wins as soon as the other player has less than 1 hit point. Damage can be done through special powers or through the cards themselves. For example, played minions (also having hit points and attack values) can directly attack the other player or minions of that player, allowing for tactics. One turn is limited to 75 seconds but can be ended earlier by the player⁴⁸. This game was also suitable, as it offers a (real-time) log file containing information on all game actions done. Thus, this log can be used programmatically (i.e., we can utilize this in a system concept). Furthermore, as personal experiences showed, many streamers of that game already discuss their play options with their audience. Therefore, by enhancing the communication and interactive channels around this, it will most likely not feel artificial.

⁴⁶ GitHyp: *Twitch's Most Watched Games of 2016*,
<https://goo.gl/APcA2q> (last accessed: 2018-07-07)

⁴⁷ Twitch Metrics: *The Most Watched Games on Twitch*, July 2018,
<https://goo.gl/qke8hf> (last accessed: 2018-07-07)

⁴⁸ All game rules can be found here: <https://goo.gl/TJX44m> (last accessed: 2018-07-07)

Based on this game, we created *Helpstone*. It has the goal to provide more sophisticated communication and interaction channels in which game-related feedback and hints can easily be given by the viewers and are presented in an easy-to-analyze way for the streamer. Viewers have access to a web page embedding the *Hearthstone* live-stream and further functions. Although we saw in the previous online study that this is not optimal, we deemed it necessary as at the time we developed *Helpstone*, the options to customize live-streaming pages were fairly limited⁴⁹. Figure 5.4 shows the resulting web page; the different aspects will be explained below. The second part is the streamer component. For this, we used the *Overwolf*⁵⁰ plug-in system, an HTML-based framework which can be used directly in game windows and allows for interactions. With this, the streamer can simply play as he or she is used to, but receive additional benefit through the overlays. Figure 5.5 shows an overview of *Hearthstone* with the overlays (which will also be explained below). The following live-streaming issues have been considered in the design of *Helpstone*:

Lag: To ensure that all information presented to the viewers relates to the in-game situation shown by the live-stream (see lag issue explanation in Section 2.5.1), users can compare and adjust a clock on the web page with a clock live-streamed by the streamer (and thus shown in their video feed). In this way, *Helpstone* widgets on the viewers' web page are synchronized with the live-stream. We can thus ensure that viewers are not confronted with information that they have not yet seen in the stream itself, to minimize confusion. For this, we utilize real-time game state information, parsed from the *Hearthstone* log.

Information overload: For the different features of *Helpstone*, we follow the concept of *ballot box communication* [330], i.e., we limit the options viewers have and aggregate these inputs. Only the aggregations are then visualized, to reduce information overload (for both streamers and viewers). To make individual contributions matter, though, we also consider mechanisms such as upvoting, to, in theory, allow others to spot good contributions faster.

New viewers: Viewers can join a channel at any time and might have missed important elements in the game; hence, they need on-demand information about the current game state [26] to make well-informed inputs. Overall, this will also minimize unintended "trolling" (i.e., without knowing the current state, suggestions of new viewers could be perceived as "trolling"). For this, some of the *Helpstone* features also provide historical information.

Enhanced interaction and communication channels: Both the streamer's and viewer's views consist of different widgets that provide the enhanced interaction and communication channels of *Helpstone*. These will be more specifically explained in this section. While Figure 5.4 and Figure 5.5 show the complete view, Figure 5.6 and Figure 5.7 show some of the widgets enlarged:

⁴⁹ More options are available today (see Section 2.5.1), but for example upvoting chat messages directly in the platform's chat is still not easy to realize with these.

⁵⁰ <http://www.overwolf.com> (last accessed: 2018-07-07)

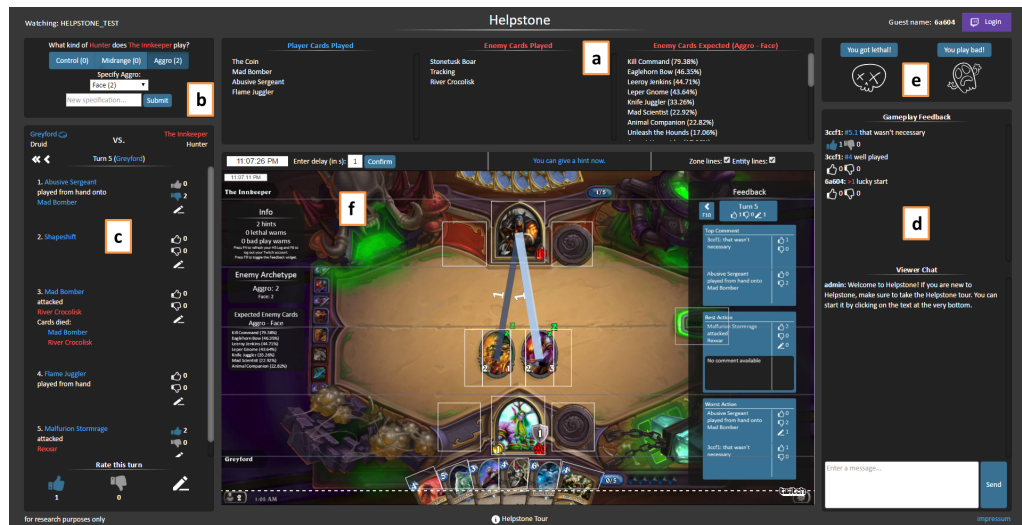


Figure 5.4: *Helpstone* from a viewer's view: a) Cards Tracker. b) Archetype Voting. c) History. d) Chat and comments at the top. e) Emergency Buttons. f) The streamed video with the Stream Overlay.



Figure 5.5: *Helpstone* from a streamer's view: g) *Hearthstone* game with the Stream Overlay. h) Statistics showing the Emergency Button presses as well as the amount of hints given. i) Archetype Voting result and Expected Cards. j) History overview of the streamer.

Stream Overlay: To simulate natural game interaction for the viewers and get sophisticated visualization of hints, every viewer can draw lines on the video stream. These lines are directly shown to the streamer via a transparent full-screen *Overwolf* widget (see Figure 5.5g). For the viewers, areas of interest are visualized

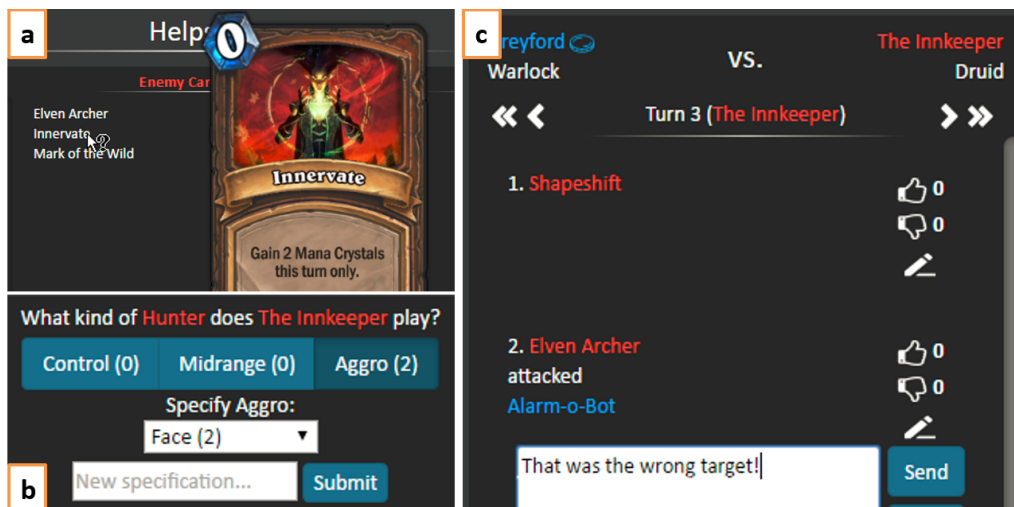


Figure 5.6: Enlarged widgets (see Figure 5.4): a) Shows a part of the Cards Tracker, with a mouse-over event. b) Archetype Voting. c) History.

by rectangles (see Figure 5.4f); only here, lines can start or end. Drawing lines is a similar interaction as directly interacting with the game *Hearthstone* (here, cards are also played by the player by dragging them to the designated spot). When in the same zones, lines of different viewers are aggregated by *Helpstone* and become thicker to make popular moves visible at a glance for the streamer. Whenever the streamer performs an in-game action, the lines are cleared to allow the audience to provide new suggestions. This widget serves as the main interaction channel, as now, hints to specific moves in the game do not need to be entered into the chat by viewers (where they would not be aggregated at all) and are directly visible to the streamer.

Cards Tracker: For historical information, played cards (from the streamer as well as the enemy) are tracked by extracting this information from the *Hearthstone* log, allowing the viewers (including ones who joined later) to get an overview of the current match. With a mouse-over, the original game cards with all details are shown (see Figure 5.6a). Together with the Archetype Voting (see below), this widget helps viewers to make informed decisions. We also show card predictions, based on past games and previously seen cards, for the most-voted-on archetype/specification. The streamer (see Figure 5.5i) can also see the card predictions inside the game. We integrated these predictions to spark more discussion around the viewers and streamer.

Archetype Voting: To speculate about the enemy’s strategy together with the streamer, the opponent’s play style (what we call “archetype”) can be classified (similar to openings in the game *chess*). Viewers can vote on the three main existing play styles. As many card deck variants exist for every archetype, we also allow viewers to vote on a specification (see Figure 5.6b). The most-voted-on specification and the number of votes for every archetype are shown to the streamer (see Figure 5.5i).

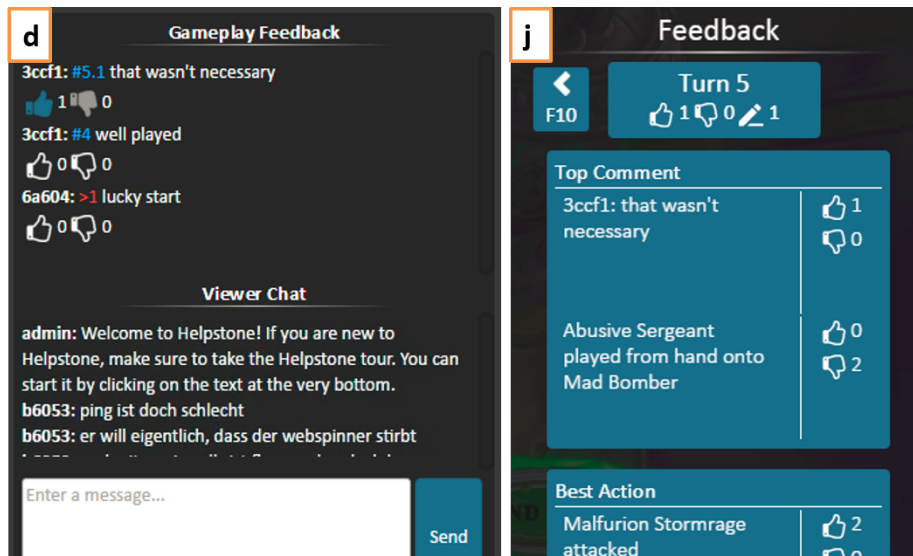


Figure 5.7: Enlarged widgets (see Figure 5.4 and Figure 5.5): d) Chat with the History comments at the top. j) Part of the History for the streamer.

History and Chat: For historical information, all turns, with details on every action and involved cards (see Figure 5.6c), are accessible to the viewers. Similar to the Cards Tracker, the cards can be inspected via mouse-over. Viewers can rate actions and turns (“thumbs up” and “thumbs down”) and can also provide a comment. These are then visualized in a dedicated chat area (see Figure 5.7d), in which comments can also be up- or downvoted and are sorted accordingly. Besides these specific elements, the chat works similarly to the *Twitch* chat. The streamer can toggle a specific overview directly in the game (see Figure 5.7j): the best-rated comment is then shown, and the best-/worst-rated action in this turn, as well as the comment that was rated best for these actions. This should allow streamers to discuss the important points with their audience in a structured way.

Emergency Buttons: As quick and easy-to-analyze feedback, viewers have the option to indicate, by a simple button press (see Figure 5.4e), that the current situation can be considered as a “bad play” or that the streamer could win the game (“has lethal”). The streamer always sees the number of bad play and lethal hints for his current turn (see Figure 5.5j). When a threshold is reached the widget starts to blink red for the streamer.

When a group of people provides similar inputs (for example, move suggestions or upvotes), these are highlighted through the system by visual means (for example, move suggestions are aggregated, lines become thicker and upvotes set comments higher in the list). While this should reduce the information overload (as explained above), it also relates to the *wisdom of crowds* [289] (see also Section 2.4) and thus provides the streamers with a range of options they can consider taking, or simply talk about them with the viewers. Either way, *Helpstone* should thus allow viewers to impact the experience.

5.3.2 User Study with *Helpstone*

Following **Goal_{HS} 2**, *Helpstone* was evaluated in an exploratory way with the goal to receive insights on how viewers perceive the enhanced interaction channels.

Method

We recruited a German streamer (25-year-old experienced male *Hearthstone* player, streaming since November 2015) who usually has 20 to 50 viewers. Additionally, we recruited people who know/play *Hearthstone* via student mailing lists and *Facebook*. Prior to the streaming session, we explained the system to the streamer. At the beginning of the stream, a four-minute-long tutorial video was shown in the live-stream to provide the viewers with a tutorial on *Helpstone*. Additionally, an integrated HTML tour on the web page explained all elements for viewers who joined later. The streamer played against a simple computer opponent to let the viewers and himself get familiar with the system. Thereafter, the streamer played against a strong computer opponent, and one round against one of his viewers. When playing against a computer opponent, *Hearthstone* does not restrict the turn time, which was beneficial for the streamer and viewers to experience the setting without time pressure. The viewers were encouraged to use *Helpstone*; all interactions were logged.

After these matches, the stream was ended and a link to an online questionnaire consisting of demographic questions, the *System Usability Scale* (SUS) [29] (to measure the overall usability of *Helpstone*) and statements on the use and the perception of every widget was provided. Respondents had to express their agreement with these statements on a 5-point scale with the labels *disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, *agree*. These statements were introduced to make the widgets comparable and receive specific insights into each of them. Free-text answers were also allowed for every widget and for the overall system. The streamer was interviewed in a semi-structured way by two researchers to receive insights into his perception of the system and options for the audience.

Results

23 viewers visited the *Helpstone* website during the study. Ten (eight male, two female; age: <18: 1, 18–24: 4, 25–30: 5) completed the questionnaire. Half of them were regular viewers and nine respondents reported being at least moderately skilled in playing *Hearthstone*. Five of the participants reported watching game live-streams nearly every day, one watches multiple times a week, two once a week and three a couple of times in a month. We asked the sample whether they have an issue with visiting another page for using *Helpstone*, but they tended to disagree to this statement ($M=2.1$, $SD=.9$, $Mdn=2$). Thus the sample can be considered as suitable for this study, and the following aspects were found:

Element	1 st match	2 nd match	Total
Drawing lines	61 (16)	43 (12)	104 (20)
Player up-/downvotes	50 (9)	23 (7)	73 (12)
Opponent up-/downvotes	20 (7)	10 (5)	30 (12)
Bad play warnings	11 (5)	7 (7)	18 (10)
Comments on player action/turns	8 (5)	10 (6)	18 (8)
Archetype votes	-	10 (10)	10 (10)
Up-/downvotes on comments	2 (2)	11 (5)	13 (7)
Comments on opponent action/turns	2 (2)	1 (1)	3 (3)
Lethal warnings	0 (0)	1 (1)	1 (1)
Total interactions	154 (16)	116 (13)	270 (22)

Table 5.8: Number of interactions with *Helpstone* per match and element. The number of unique users using it is provided in parentheses. Note: The Archetype Voting is not supported against computer opponents. The 1st match took 14 and the 2nd 12 minutes.

***Helpstone* subjectively raises the audience activity level and perceived influence:** The viewers enjoyed *Helpstone* ($M=4.4$, $SD=.5$, $Mdn=4$) and by using it reported being more active while watching the live-stream ($M=4.5$, $SD=.7$, $Mdn=5$). The viewers tended to agree that *Helpstone* helped them to better interact with the streamer than they could over the *Twitch* chat ($M=3.8$, $SD=1.2$, $Mdn=4$), but were indifferent whether this was also true for the interaction with the other viewers over the *Twitch* chat ($M=2.8$, $SD=1.6$, $Mdn=3.5$). The latter aspect can be attributed to the fact that we did not integrate the *Twitch* chat directly (i.e., chats entered on the channel's *Twitch* page were not shown in *Helpstone* and vice versa). This was deemed as not necessary for the study but can easily be integrated. The viewers had the feeling of exerting influence on the streamer ($M=4.5$, $SD=.5$, $Mdn=4.5$) and the streamer also reported in the interview the feeling that he had been influenced, even though he played in "his style". Nonetheless, unsurprisingly, the sample also tended to agree that using *Helpstone* takes more effort than watching a "normal" live-stream ($M=3.9$, $SD=.9$, $Mdn=4$).

***Helpstone* increases game-related interactions:** We tracked the viewers' activity on this channel before our study, for three consecutive *Hearthstone* matches (which took 12 minutes in total), and found that 30 to 40 viewers were watching. Out of those, seven wrote 22 chat messages; none of them was game-related. A later tracking of one-hour footage of this channel with a similar audience size showed that only 18 of 144 chat messages (12.5%) were game-related. In the interview, the streamer stated that on his channel more social than game-related conversations happen. Further, he states that his viewership indeed provides hints – especially when he plays badly – and that he sometimes asks his viewers game-related questions, making *Helpstone* reasonable for his channel. The two matches in our study lasted 26 minutes in total, revealing that *Helpstone* will prolong matches, to give the audience room for suggestions. Table 5.8 shows the

I think it is reasonable...	M	SD	Mdn
... to suggest actions to the streamer (by drawing lines)	4.5	.5	4.5
... that viewers can vote for comments ("thumbs up"/"thumbs down")	4.2	1	5
... that viewers can rate actions ("thumbs up"/"thumbs down")	4.2	1	4.5
... to be able to warn the streamer of lethal situations	4.1	1	4
... that viewers can comment on actions	4.1	1.3	4.5
... that viewers can suggest new specifications	3.9	1	4
... that viewers can comment on turns	3.9	1.6	4.5
... to see a prognosis of the opponent's remaining cards	3.7	1.2	4
... that viewers can rate turns	3.7	1.5	4
... to be able to vote for archetypes	3.4	1.3	4
... to be able to warn the streamer of urgent bad play	2.9	1.4	2.5

Table 5.9: Relevance statements for the different *Helpstone* elements.

Widget	Aware			Usage			Agg			SAware		
	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn	M	SD	Mdn
Stream Overlay	4.7	.5	5	4.5	.5	4.5	3.9	1.2	4	4.7	.5	5
Cards Tracker	3.4	1.1	4	2.7	1.1	2.5	-	-	-	-	-	-
Archetype Voting	3.3	1.3	3.5	3.3	1.3	3.5	3.8	.9	4	4.4	.7	4.5
History	4.2	1.2	4.5	4.1	1.3	4.5	4.2	.6	4	3.9	1.1	4
Chat	4.1	1	4	4	1.2	4	-	-	-	4.4	.5	4
Emergency Buttons	4.2	.9	4	3.8	1.3	4	3.9	1.3	4	3.8	.8	4

Table 5.10: Mean values, standard deviations and median values for the following statements: Aware: *I think that I actually perceived the widget*; Usage: *I think that I used this widget actively*; Agg: *I liked the aggregation of viewer inputs for this widget*; SAware: *I think that inputs for this widget were noticed by the streamer during a match*.

number of interactions with every *Helpstone* element and indicates that through *Helpstone*, more game-related interactions happen. The artificial situation and the novelty effect [107, 148, 332] might have had a strong influence, but the numbers of drawn lines and the upvotes (the two most often used features) especially indicate that viewers' game-related interactions might increase if a tool enables more interaction and communication channels.

***Helpstone* offers relevant features, even for passive viewers:** Table 5.9 shows that not all new features were perceived as equally relevant; thus, not every option for more interaction should be offered unconditionally. Further research needs to be done to investigate whether, in a stream, interaction options, that are not perceived as relevant are harmful. Overall, this finding fits with the differences in the feature perception found in the online study (see Section 5.2.2). We also asked participants about several statements regarding the features of *Helpstone*; these can be found in Table 5.10. Interestingly, we see that the Cards

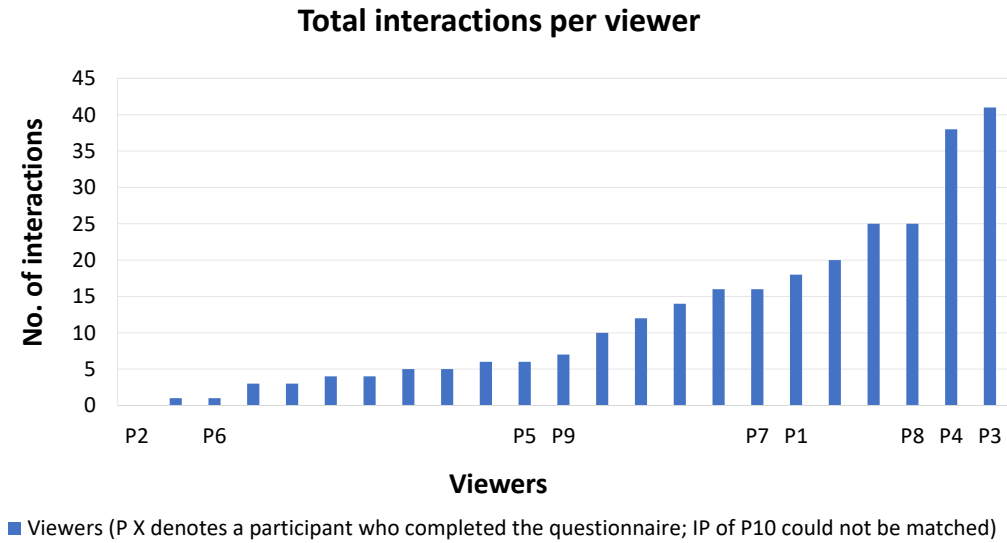


Figure 5.8: Number of interactions done with *Helpstone* per viewer.

Tracker and the Archetype Voting were not noticed as much by participants as the other widgets, which also explains why the voting for archetypes was less relevant, for example. By considering Table 5.8 and Table 5.9, we conclude that the stream overlay was perceived as the most important element, even though usability flaws were reported (see next result). It seems that the direct interaction with the video stream and the immediate feedback are promising for such systems. This also directly fits to the recent platform efforts to make streams more interactive (see Section 2.5.1). Rating game actions (see Table 5.9) and presenting the results in an aggregated fashion was appreciated by the viewers ($M=4.2$, $SD=.6$, $Mdn=4$) and explicitly approved by the streamer. This aspect can also be seen in Table 5.10, as the aggregation functions were rated above average. We conclude that interactive features which provide more discussion options for the streamer are important. This also adds to the aspect that streamers and their performance are very important for viewers in a stream (as also found in the online study; see Section 5.2.2). Finally, it seems that for functions providing information on the game state, easy-to-verify ground-truth information (lethal warnings) is preferred over subjective or ambiguous hints (bad play warnings).

Considering the interaction count of the viewers who provided answers to our questionnaire (see Figure 5.8), we see that some only did a few interactions. The assessment of the different features in Table 5.9 shows that these (passive) users had also positively rated the interactive features; otherwise, the mean values would have been worse (and furthermore, t-tests between viewers with low (≤ 10) and high (> 10) numbers of interactions did not reveal any significant differences for the different ratings). This further supports the aspect, found in the online study (see Section 5.2.2), that viewers who want to watch rather than to interact see usefulness in the new interaction channels as well.

Helpstone's usability can be further improved: The streaming lag in our case was not an issue, as participants tended to disagree to the statement that the stream delay was annoying ($M=2.3$, $SD=1.1$, $Mdn=2.5$). At the same time, the sample disagreed to the question whether it was difficult to adjust the lag on *Helpstone* ($M=1.5$, $SD=1$, $Mdn=1$). Taken together, this indicates that the lag solution we used in *Helpstone* might be utilized for other (round-based) games as well. Considering usability, the *SUS* score was $M=70.3$ ($SD=13.5$). According to [10], this is acceptable. Two major usability issues were reported: first, the History always advances to the next turn, which sometimes causes viewers to rate or comment on the wrong turn. Second, the Stream Overlay needs refinement to be more suitable for *Hearthstone* matches (to suggest chains of actions). The streamer also reported that the Stream Overlay interferes with in-game information (e.g., card texts). However, both widgets were seen as relevant (see Table 5.9) and were used objectively (see Table 5.8) and subjectively (see Table 5.10). Overall, while in its current form the sample is indifferent as to whether *Helpstone* should be used for all *Hearthstone* streams ($M=2.8$, $SD=.9$, $Mdn=2.5$), after improvement (i.e., integrating the usability feedback and suggestions made) participants would be more open to this ($M=3.9$, $SD=1$, $Mdn=4$). This provides further support for the view that having more interaction channels in live-streams appears to be appreciated overall.

Discussion

The case study revealed that *Helpstone* indeed provides enhanced communication and interaction channels, as these were rated positively and used by the participants in the study. It gives the audience a feeling of influence, and the streamer reported to also have the impression of being influenced by the audience. Additionally, *Helpstone* raised the audience activity level, and may also be interesting for passive viewers who simply want to watch the stream.

To our knowledge, this was the first investigation of such a setup that allows a better direct interaction between viewer and streamer in the live-streaming context. However, the results should not be over-generalized, as we had only a small sample size, only one streamer, and we focused only on a round-based game. We deem these limitations acceptable for a first exploration of this topic. Not every function offered seems equally relevant to viewers. It seems necessary to have rules to know what works in which situations. In this sense, this also relates back to the online study (see Section 5.2.2) that showed that there are contextual and individual differences. Further research can build upon our concepts and results, especially by using *ballot box communication* and direct interactions on the video stream, as these seem promising for novel interactive streaming tools. It can easily be seen that the chosen concept is applicable to other turn-based games with a fixed camera as well, such as *chess* or *poker*. Continuous game play and a moving camera add complexity, but even here, parts of our concepts can be used: for example, for first person shooters, comments and

upvotes could show which weapons to use next, while for role-playing games, dialogue options to be selected by the streamer could be made interactive for the viewers on the video stream. Work on such tools could create completely novel experiences for streamers and viewers.

5.3.3 Contribution to the Thesis' Questions

With this approach, we could show that providing viewers with more influence in live-streams is worthwhile. With *Helpstone* we demonstrated how enhanced communication and interaction channels can be realized (**Goal_{HS} 1**) to grant viewers more influence on how the stream proceeds. *Helpstone* also provided solutions to current issues imposed by the live-streaming platforms, such as a workaround for the lag issue in respect to the available information (see Section 2.5.1), information overload and the on-boarding of new viewers. The interaction patterns *Helpstone* provides also allow viewers to stay passive and simply consume the new functions (e.g., such as the history), which we learned is indeed reasonable, further strengthening the results of the online study (see Section 5.2.2). Through the conducted study (**Goal_{HS} 2**), we also found that the group of viewers can indeed exert influence on the streamer, as he considered their suggestions. In addition, they exert influence directly on the stream, as their contributions, while orchestrated by the streamer, were also visible on the stream when he decided to use them. Considering that live-streaming has the goal to be more interactive than simply watching a video (as discussed in Section 1.2.2 and Section 2.5), and as viewers reported that they could imagine always using *Helpstone* (after adding their suggestions) in streams, we assessed the means that *Helpstone* offers as worthwhile to raise the individual autonomy. Overall, this fits with the questions of thesis (**RQ3**; see Section 1.4) as we demonstrated how enhanced interaction channels can be realized and that these are indeed perceived positively.

5.4 Summary

This chapter considered game live-streams in which many-to-many (viewers amongst them) and many-to-one (viewers to streamer) communications are relevant. The video analysis of a large channel that seeks to integrate viewers showed by what means they are empowered today. We found that there are several, but that viewers do not have unrestricted influence. Instead, the influence is always moderated. In the complementary online study, we found that this is also considered as valuable by viewers, as they do not want the streamer's performance to be affected unconditionally. At the same time, though, we also found that many highly rated features, concepts and streamer's behaviors are interactive in nature. This underlines that game live-streams are a relevant context to investigate how to empower individual users. This study revealed further aspects in relation to integrative options, such as that viewers who do

not want to participate still assess interactivity as reasonable. Finally, with *Helpstone* we demonstrated how to create a system that aims at empowering users, and we could validate that the new interaction options for individuals are indeed perceived as valuable. Taken together, we found that it is interesting for individuals to have more autonomy in the live-streaming context, but only within limits, as unrestricted autonomy could lead to a negative experience for everyone based on the shared medium. Instead, here, the streamer should orchestrate the contributions. The findings in this chapter add to **RQ3** (see Section 1.4) and to the ongoing efforts of empowering viewers in live-streaming settings.

The next chapter will analyze how viewers interact when there is no streamer to orchestrate their actions. For this we consider settings in which the audience alone is responsible for what is happening in a streaming context. We will consider shared game control and investigate what means can be provided to allow a group of people to manage itself.

Chapter 6

Shared Game Control in Live-Streams

This chapter presents work that we conducted in shared game control settings. We investigate these settings in the live-streaming context with two shared game control prototypes, *TPP++* and *CrowdChess*. This fits into our research questions (and adds especially to **RQ3**; see Section 1.4). Within this thesis, in contrast to Chapter 5, it allows us to study group settings, without orchestration of a streamer, as the latter is absent in such scenarios. Thus, this chapter shows how the audience is able to self-orchestrate and how they play such games together. With the conducted studies, we are able to reason about the viewer’s perception of such settings and how effective the group decisions are.

Sections 6.2 and 6.4 are based on the publication [171] and Section 6.3 is based on [172].

6.1 Introduction

In this chapter, we consider shared game control settings in live-streaming. As illustrated in Section 2.5.2, *Twitch Plays Pokémon (TPP)* was a particularly successful representative attracting many people to play in parallel, and many other such settings appeared subsequently. For this thesis, especially for **RQ3** (see Section 1.4), such settings are particularly interesting. While Chapter 5 has considered the question of how viewers can be empowered in streaming situations in which a streamer is present and able to orchestrate the contributions, in this chapter, we will investigate how viewers orchestrate themselves: in these games, individuals can shape what happens in the game, but as this is a shared experience, the viewers need to manage themselves and thus can solely decide on how the stream proceeds. While we described self-administration options in Section 2.5.2 (e.g., the “*anarchy*” and “*democracy mode*” in *TPP*), we were interested

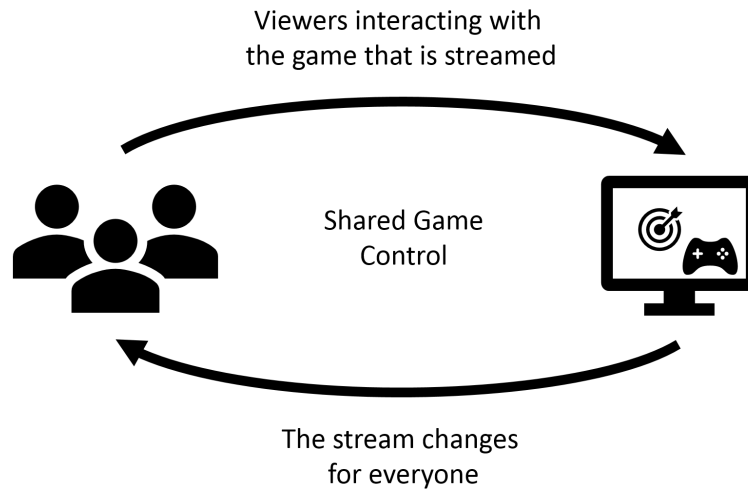


Figure 6.1: Instantiated schematic of reciprocity in shared game control.

in providing further options and learning about the effectiveness of the audience in playing together. In the self-sustaining systems presented in Chapter 3, users were only loosely coupled. Here, in the shared game control context, users can be considered as tightly coupled. They share the same view (i.e., the outcome directly affects the stream for everyone) instead of each having their own “view” in which individual contributions might change only small bits for other users (i.e., when part of a receipt in *ExpenseControl* is corrected). Figure 6.1 shows an instantiated schematic of the reciprocity in shared game control.

This chapter is structured as follows: we introduce an extended *Twitch Plays Pokémon* setting, *TPP++*. Here, viewers received more input aggregators beyond the “anarchy mode” and “democracy mode” and access to game elements. Both aimed at providing more self-administration options. After presenting the system, we elaborate on a small-scale study we had conducted with it, showing the relevancy for both means. We will then elaborate on *CrowdChess*, a shared game control variant of *chess* to measure the effectiveness of individual and group decisions and a study we conducted on the live-streaming platform *Mixer* with it.

6.2 Perception of a Shared Game Control Setting

Based on the work presented on *TPP* in Section 2.5.2, we were interested in whether similar channels are equally attractive. To this end, we considered ten days of viewer count data in April 2016 from streams that appeared on the *Twitch Plays*⁵¹ section on *Twitch*, which provides an overview on *TP* channels. Every 20 minutes, we checked the viewer count in channels listed there. 58 channels appeared, and the massive numbers seen during the first instance of *TPP* were

⁵¹ We will abbreviate live-streaming-related shared game control with *TP* subsequently, independent of on which of the live-streaming platform they are presented.

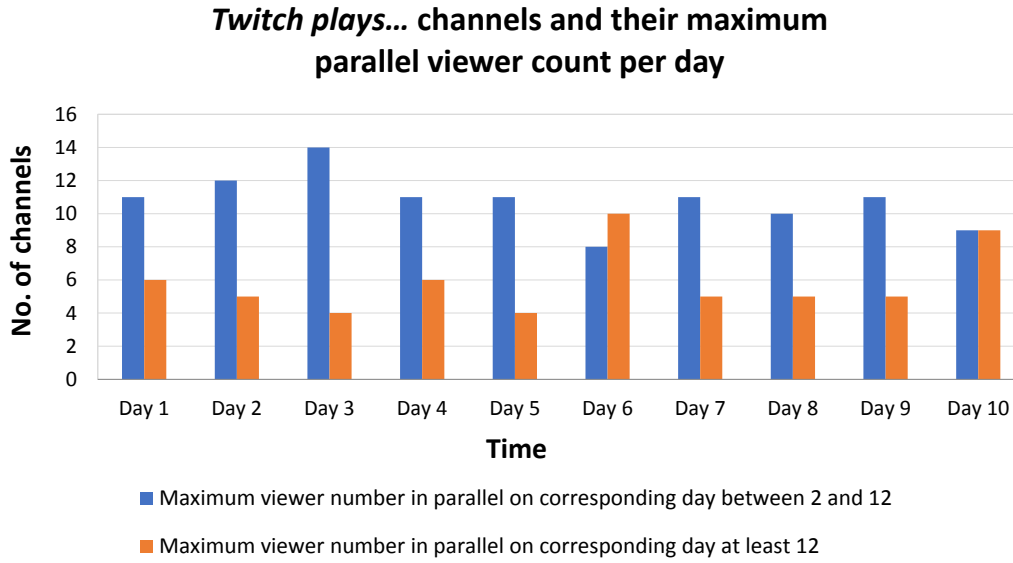


Figure 6.2: Maximum parallel viewer count per day for *TP* channels.

not reached by *TPP* itself or any other *TP* streams. The channel with the most viewers in parallel had 7689 viewers on one day and was a notable exception. *TPP* had 322 viewers in parallel at the peak. Apparently, there are two classes of channels (see Figure 6.2): a few channels that attract a larger number of viewers, and channels that have a small viewership in parallel (<13). Thus, apparently the *TP* channels also have a long-tail distribution, similar to the usual streamer-driven channels in live-streaming [139].

Based on this finding, we assessed it as interesting to start the investigation of the shared game control setting in live-streams, by considering how viewers perceive such an experience. Thus, we had the following goals:

- Goal_{TPP} 1** *Creation of a shared game control setting with enhanced self-administration options:* Following the idea of *TPP*, we were interested in creating a *TP* system that offers more options than the original *TPP* for individuals to impact the experience and allow the group to potentially better self-administrate. We judged the enhanced options as necessary to give the individual viewers more influence, and potentially also increase the perception of such an experience.
- Goal_{TPP} 2** *Evaluation of the self-administration options in, and perception of, a shared game control setting:* We were interested in how *TP* experiences are perceived by viewers and to what extent the audience uses the added self-administration options.

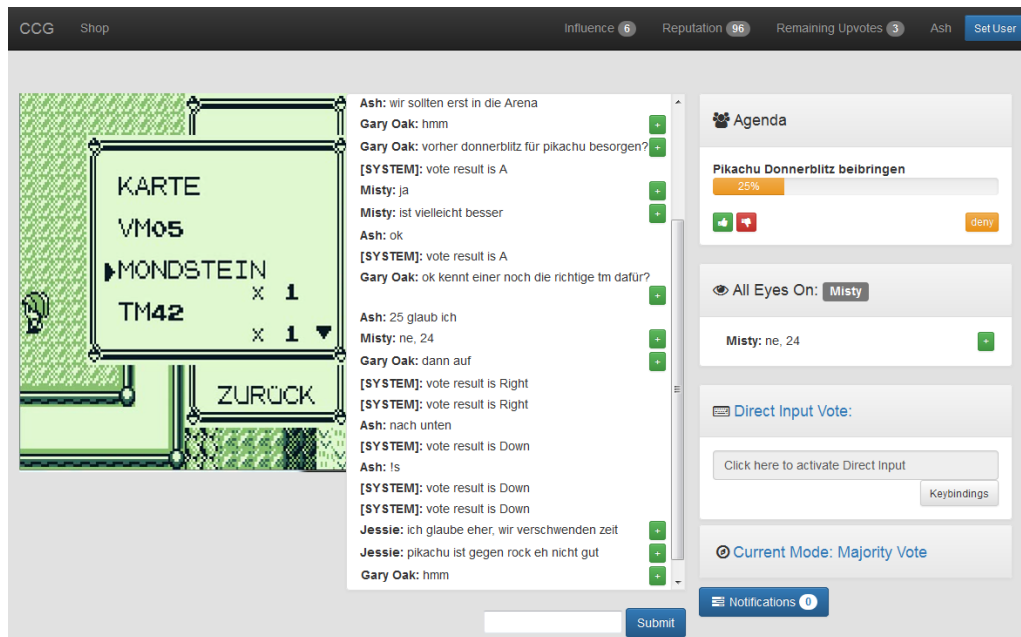


Figure 6.3: The TPP++ (web-based) interface.

6.2.1 Concept and System Design of TPP++

Following **Goal_{TPP} 1**, we created TPP++ (see Figure 6.3), based on the original TPP. We developed a server and a web application to manipulate the setup more easily than could be done by using *Twitch*, and also to mitigate the lag issue (see Section 2.5.1). As a basis, we adapted elements used on *Twitch*, i.e., the login, chat and streaming window, as well as the option to enter commands into the chat that are interpreted in the game. We added an option to directly use keystrokes to carry out commands (instead of using the chat). As the “*democracy mode*” in TPP was introduced to make progress faster [187] (see also Section 2.5.2), we were interested in exploring further modes. For these we were inspired by work on crowd input aggregation (see Section 2.4.3). Similar to TPP, the audience has the option to alter the aggregator used, to account for their needs in the game. Besides the aggregators, we added further means for self-administration that can be used complementarily. We framed them as game elements. The online study (see Section 5.2.2) showed that game elements seem unnecessary in the presence of a streamer. However, in the absence of one, this might further add to the experience. Both the aggregators and game elements have the goal to raise engagement and the level of audience self-administration.

Aggregators: For the aggregators (see Table 6.1) we provided a mode in which all commands are executed (Mob), one in which non-experts have more weight (Proletarian), a mode in which the audience can give certain viewers more influence (Expertise Weighted Vote), modes in which the best conforming decision is selected (Majority Vote, Crowd Weighted Vote) and modes in which an individual

Aggregator name	Aggregator functionality
Mob	<i>"Anarchy mode"</i> in TPP and Mob/Multi in [160]. Every vote is processed and carried out.
Majority Vote	<i>"Democracy mode"</i> in TPP. Time frames are considered and the most-selected command is used.
Crowd Weighted Vote	Based on [160, 260]. It is a weighted majority vote, where the player's weight is based on their conformity to the others. The weight of all players is continuously adapted based on their voting and the most popular vote, i.e. increased (decreased) if the choice is (not) congruent with the most popular vote.
Active	Based on [160]. If this aggregator is chosen, one player is randomly selected and takes control over the game, as long as input from him or her is provided or the aggregator is changed.
Leader	Based on Legion Leader [160]; combines Active and Crowd Weighted Vote, i.e., a player is selected based on his or her conformity to the crowd.
Expertise Weighted Vote	A weighted majority vote with weights based on the expertise of the players (often used in crowd-based systems [239]). In our case, the individual expertise is generated by the players themselves, as players can receive upvotes (which improves their reputation; see the game elements section) for their actions, i.e., the players can identify others they want to provide with more weight in decisions.
Proletarian	The inverse of the Expertise Weighted Vote, to now empower non-experts.

Table 6.1: The aggregators (and explanations thereof) offered in TPP++.

decides (Active, Leader). Every player can change the desired mode, but only the most often selected becomes active, and this is visualized on the web page. With this set of aggregators we offered more choices than the original TPP provided and in particular allowed the audience to decide whether they want to empower individuals for certain parts of the game.

Game elements: With these, we introduced an orthogonal option to the aggregators for orchestration. We use two values, called *influence* and *reputation*. While reputation is a permanent value (used in Expertise Weighted Vote), influence is used for buying items. Every player can upvote others, for example after they have provided good suggestions. The amount of upvotes a player can spend is limited, but refreshes over time. Every upvote generates influence and reputation for the upvoted player. Influence can be spent on these items:

- **Agenda:** A player-generated short-term goal (e.g., *"Go to city X"*) can be established, which might be beneficial for motivation [177] and furthermore might help to orchestrate the group. It is shown to the audience and they can vote on whether the goal has been reached or not, or state that they do not want to pursue this goal. If a decision has reached a majority (among all logged-in players), or the majority has cast their vote, the agenda ends. Participating players gain a small amount of influence independent from the outcome, in order to prevent incentives to manipulate agenda votes for influence gain. Only one agenda can be active and no agenda items can be bought during this time.

- **Player Spotlight:** One detrimental factor in user motivation within a crowd is the invisibility of one's own contributions [14]. This item tries to motivate players by highlighting their actions for all others. The featured person is chosen randomly and showcased for a fixed period of time, after which the next one is selected. Players can cumulatively increase their chance of being chosen by buying this item.
- **Repay:** This item distributes its influence cost evenly among peer players. Loparev et al. [178] introduced a passive game mode which allowed players to aid their fellow players without directly influencing the game state. Additionally, in cases of a skewed distribution of influence, which for example might occur when one or a few players are favored by others due to their expertise, this item can be used to re-balance.

6.2.2 User Study with *TPP++*

TPP++ was used in a user study to learn how the aggregators and the game elements are perceived and used by players and what requirements and expectations users have towards them. In this sense, this study was an exploratory one, not driven by hypotheses. Nonetheless, we expected that the introduced means will help the group to better self-administrate. With this study, we were also interested in assessing the overall perception of a *TPP*-like setting (**Goal_{TPP} 2**). It can be expected that the dynamics in such streams differ, the more viewers take part. As it was shown that most *TPP*-like channels attract only a smaller viewership, we also aimed for a smaller group of viewers in this study.

Method

We used *TPP++* in a local setup to minimize streaming delay even further. All participants were physically separated. In addition, a list of user names was handed out, and it was forbidden for the players to reveal their identity. Both were meant to mimic the *TPP* setting, in which participants probably did not know each other. After a pre-session questionnaire (assessing demographics and subjective experience with computer games and *TP* settings), every participant received access to a *Pokémon Red* game and had ten minutes to get familiar with it. An interactive explanation of how the aggregators work followed: the participants were able to define voting options and values, and could see what a selected aggregator would output. The user interface of the web page was also explained. Afterward they played the game together, similar to *TPP*. This part was separated into four phases, representing different situations and options: the first two phases did not use further game elements, restricting the user interface to the login button, stream window, chat, aggregators (without Expertise Weighted Vote and Proletarian) and direct input option. In order to avoid demotivation, the

phases without game elements had to happen before these were introduced, since taking away features could have a negative impact on the user experience [300]. The four phases were (see Section 2.5.2 for an explanation of the game areas/*TPP* situations mentioned below):

- **Easy, no game elements (ENG):** This phase started in an area in which navigation and fights are easy. Participants should get familiar with the system and the group dynamics.
- **Difficult, no game elements (DNG):** This phase starts in a difficult game area (*the Rock Tunnel*) since the screen will turn almost black, only showing silhouettes of the walls. Using a special ability of the avatar, players can illuminate their surroundings. Combined with a more challenging navigation task, this situation demands coordination from the players.
- **Easy, with game elements (EWG):** Reusing or resetting known scenarios could frustrate players who have their achieved progress reset; thus, for the game element conditions we needed to ensure the use of other states fulfilling the requirements. Thus, we selected a similar but different state compared to ENG, i.e., easy navigation and easy fights.
- **Difficult, with game elements (DWG):** The phase started in the difficult game area called *Spinning Hell*. A high amount of floor tiles move (spin) the character in different directions, making navigation through the maze hard. This area led to the introduction of the “*democracy mode*” in *TPP*, because progress in “*anarchy mode*” proved impossible [2].

Every phase was active for 15 minutes of game time. At the beginning of each phase, participants receive information about the available creatures, items, special abilities and the story state. No further requirements regarding method of play were given, i.e., after they received the overview, participants could play the game as they wanted. Between each phase they were asked to answer questions regarding their enjoyment, perceived progress, difficulty and usefulness of the available features on a 7-point scale (with labels at the extreme values and the midpoint: *strongly disagree*, *neither agree or disagree*, *strongly agree*). After the last phase, the provided questionnaire contained an additional section including a general assessment of *TPP++*. A post-session interview with every participant was conducted to gather further qualitative feedback. Besides these qualitative measures, we also recorded all interactions with *TPP++*, as well as the game play. The chat was also recorded, but this led to no conclusive data.

Results

Eight German subjects participated (seven male, one female; age: 21–30: 8). People of this age and gender represent the largest user group on *Twitch*, according to their own statistics⁵². Six participants were students, two were employed, and all deemed themselves quite experienced with video games ($M=7.8$, $SD=2.6$, $Mdn=8.5$) on a 10-point scale. One had participated in at least one *TPP* session; six had already seen footage or heard of it.

Progress: The participants were able to solve tasks such as lighting and successfully proceeding within the *Rock Tunnel*, navigating parts of the *Spinning Hell* and winning fights against non-player characters throughout all phases, i.e., the group achieved progress. One explanation for this could be that we had (most likely) no destructive forces amongst the participants (i.e., no trolling tendencies became obvious) and a low participant count, in contrast to *TPP*, making it potentially easier to achieve goals. Nonetheless, from their own perception, the participants evaluated their progress as low throughout all phases: lowest in EWG ($M=2.4$, $SD=2.1$, $Mdn=1.5$) and DWG ($M=2$, $SD=1.5$, $Mdn=1.5$) and highest in DNG ($M=4.4$, $SD=1.7$, $Mdn=4.5$). As situations occurred in which advancement was slowed, mostly due to dissent on how to proceed, this could be an explanation. An example failure occurred in EWG where the group steered the character back and forth for minutes. We observed that participants with a lower overall system interaction per minute count tended to do social actions (chatting, upvoting, polls etc.) instead of issuing commands. This indicates that our audience was also not uniform (fitting into the personas [41] and results found in Section 5.2.2); i.e., different roles might be available in the audience, altering how they interact in shared game control settings.

Self-administration through aggregators: In every phase, *TPP++* started with the Majority Vote aggregator. In ENG, the first mode change occurred after 63 seconds; in the later phases the first mode change occurred after 4 seconds on average. The ability to change the aggregator was assessed differently in the phases (mean values per phase: 4/5.4/3.4/4.5) and players disagreed with the statement that all aggregators were equally important (2.1/3.3/1.7/2.6). Figure 6.4 shows the measured usage times. The most used aggregator was Expertise Weighted Vote with an overall uptime of 1028 seconds, while Mob (145s) and Proletarian (198s) were barely used. Considering the assessment of the aggregators, Crowd Weighted Vote was always mentioned as important (and players got 65 upvotes in EWG and 39 in DWG). Additionally, in the difficult phases, aggregators that provide individuals with more weight (Expertise Weighted Vote) or single user options (Active, Leader) were highlighted. While it could be argued that although Leader was chosen in DNG, all other aggregators in this phase had more uptime than Leader, a closer look at the game footage provides more insight. DNG has two major tasks: lighting the rock tunnel (once lighted it stays lighted) and navigating it. In *TPP*, the first task was never completed (they navigated

⁵²Twitch Advertising: *Audience*, <https://goo.gl/YcqXtG> (last accessed: 2018-07-07)

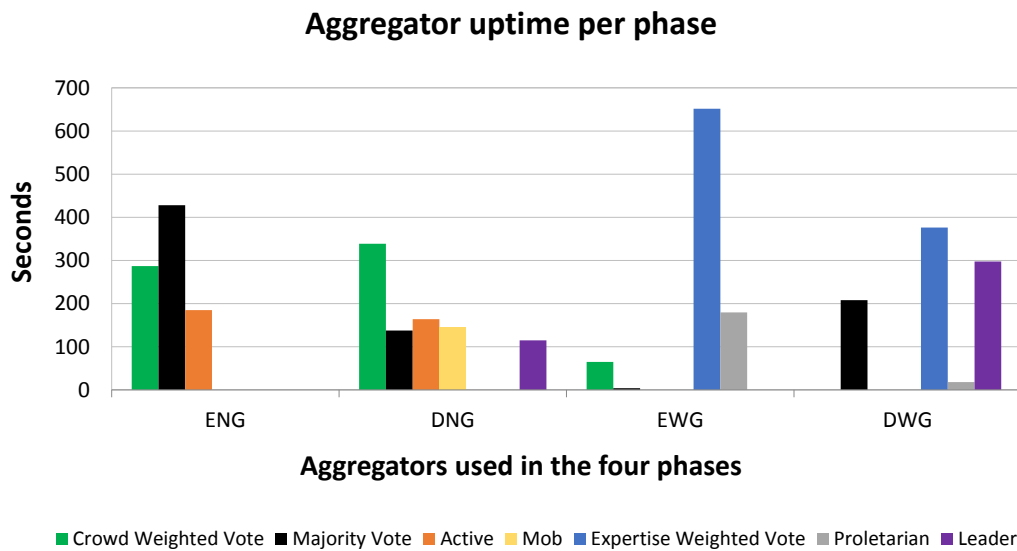


Figure 6.4: The different aggregators and their uptime in the study.

without seeing the whole map) and thus the difficulty of this scenario was quite high. In our study, the participants managed to solve the first task at the four minute mark; during this time Leader had its total uptime of 115s. Overall, the set of aggregators was rated as complete ($M=6$, $SD=.6$, $Mdn=6$) and as containing useless ones ($M=5.1$, $SD=2.1$, $Mdn=6$): Proletarian was named thrice, Mob twice and Active once, which coincides with their uptimes. Participants were indifferent (with better scores in DNG and DWG) to the statement whether aggregators provide a good option to self-administer the group (3.9/4.8/3.3/4.1) and they were indifferent (but tended to agree) to the question whether their decisions were in line with the group's decisions (4.9/4.5/4.4/5.3).

Perception and self-administration through the game elements: The sample agreed slightly with the statement that they had fun (asked on a single scale) during the study overall ($M=5.1$, $SD=1.2$, $Mdn=5$). Breaking it down to the single phases, we see that in ENG and DNG the perceived fun ($M=3.5$, $SD=1.6$, $Mdn=3.5$ / $M=3.6$, $SD=1.6$, $Mdn=3.5$) was low. This could hint that playing games that are not designed primarily for a shared game control setting needs further incentives to be fun. After adding the game elements, the self-reports for the perceived fun increased in EWG ($M=5.8$, $SD=1.7$, $Mdn=6$) and in DWG ($M=5.1$, $SD=1$, $Mdn=5.5$), even though the perceived progress was lowest in these phases, as stated. There was a significant difference in these measurements, as a Friedman ANOVA showed: $\chi^2(3)=10.9$, $p=.01$. Post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, but revealed no significant differences between the phases. When asked about whether or not game elements improved the user interface, users evaluated them as beneficial ($M=5.9$, $SD=1.8$, $Mdn=7$). They also tended to agree to the statements that these elements added to their enjoyment ($M=5.6$, $SD=1.6$, $Mdn=6$ in EWG and $M=4.6$,

SD=1.5, Mdn=5 in DWG) and were motivating (M=4.9, SD=1.5, Mdn=5.5 in EWG and M=4.5, SD=2.3, Mdn=5.5 in DWG). Additionally, these features were deemed helpful regarding self-administration in EWG (M=5, SD=2, Mdn=6) but less so in DWG (M=3.8, SD=1.8, Mdn=4), but positively influenced the group in both phases (M=5.6, SD=1.6, Mdn=6 in EWG and M=4.5, SD=2, Mdn=5.5 in DWG). Overall, 20 Player Spotlight, 15 Repay and 11 Agenda items were bought. Not surprisingly, the sample also agreed with the statement that further game elements is a good idea in this setting overall (M=5.9, SD=1.2, Mdn=6).

After every phase, we asked whether the offered functions (overall, i.e., we did not separate between aggregators and game elements) were assessed as helpful. Here, again, a slight improvement can be seen with game elements being available (4/4.1/5.9/4.6). Nonetheless, the sample tended to agree uniformly across the phases that the offered functions were not enough (4.5/4.9/5/5). An issue that was revealed in the interviews was that more shop items would be helpful and would thus expand on the self-administration options currently offered, e.g., sub-agendas or an item that provides the leader role for a short time. The latter hints that at least some participants would further empower individual control options in shared game control settings. Two participants also wished for options with more impact. These participants had a lot of upvotes and wanted to spend their influence further. The experience *TPP++* provided could also be further improved by adding more statistics, social information and the option to share responsibilities, according to the participants in the interviews. The most common demand was a static display of the current leader. Additional demands concerned seeing how much influence or reputation others have, and a visualization of how much weight the players have in each voting phase.

Discussion

To our knowledge, our study with *TPP++* was a first controlled exploration of the *TPP* phenomenon. The perceived game progress was low, even with the additional elements offered. The perceived fun was lower in the phases without game elements. Considering the content of the original *TPP* setting, the high number of players can be more likely be attributed to the novelty, the chaos, but also the “fandom” that was created around it (see Section 2.5.2). Our results indicate that simple majority polls alone are not sufficient as input aggregators here. Depending on the situation, the audience reduced the average player influence for the sake of progress and coordination, either by selecting single-user aggregators (Leader or Active) or empowering a group of individuals (Expertise Weighted Vote). This indicates that the audience does not always want everyone to contribute equally, for example to achieve faster progress. The conformity-driven aggregator Crowd Weighted Vote was selected in the first two phases but was used almost not at all later. As in later phases the Expertise Weighted Vote aggregator was available, potentially it appeared more suitable. What can be derived from this is that a fixed aggregation of individual

inputs seems insufficient. Demands vary depending on the situation, and there should be different options offered to fit these situations. Interestingly though, the audience judged the ability to change the aggregators as not particularly important throughout all phases, although changes occurred during each phase. Considering the features' ratings in the online study earlier (see Section 5.2.2), the ability to change voting schemes in polls was not rated highly either.

The aggregators in general were also not perceived as particularly useful for self-administration, even though they were actually used. First, this could be a result of integrating aggregators that were "useless" from the user view. Potentially, this could have altered the perception of the aggregators overall. Second, it could be that aggregating single-user commands is necessary to make progress, but will still lead to a different feeling than playing a game in which a user has full control. Most likely not every decision of a viewer is carried out in a *TP* setting, which might lead to such an impression overall. In contrast, considering the introduced game elements, we learned that these add to the experience and provide a better feeling of self-administration. The participants requested additional elements, not only for entertainment, but also to better influence the decisions and how the game proceeds. Integrating their suggestions and experimenting with these elements are interesting further research directions. It would also be interesting to see how these are perceived long-term.

Our study had limitations: first, although in line with the *Twitch* demographics, the laboratory study with its small and gender-biased sample. As many available *TPP* channels have a similar small audience, our study still provided actionable results. A second limitation was the short time frame the participants had to interact with the game phases. We decided to see this as acceptable, as it is currently unclear how long people interact with a *TPP*-like setting in general. 15 minutes appears to be a compromise for the participants to report their initial perception. Third, the user study itself had no players hindering progress, in contrast to the original *TPP* setting. Even though we think that the available aggregators will moderate effects that are introduced with such players, our study can draw no definite conclusions for this. Work such as [139] shows that most channels on *Twitch* attract only a few viewers; i.e., launching a channel does not automatically mean that a large viewer base can simultaneously be reached, which is why we decided against doing an "in the wild" study for this experiment. Fourth, the decision to do this study within-subject (instead of using a between-subject design/without a control group that plays through the four scenarios without game elements) could have led to learning effects (while cycling through the difficulties) and thus the impact of the game elements could have been overestimated by the participants (even though we found that less likely based on the qualitative answers). Fifth, our study used the relatively simple (considering interacting and navigating) *Pokémon Red* game. The results should be seen as applicable for this genre and not necessarily other genres, as these might demand faster interaction cycles (e.g., shooters), thereby posing different challenges.

6.2.3 Contribution to the Thesis' Questions

With *TPP++* we developed an extended version of the *TPP* setting (**Goal_{TPP} 1**). Besides offering more input aggregation modes, it also offered further means (framed as additional game elements) to allow a group to more efficiently share control over the game. By conducting a small-scale laboratory study (**Goal_{TPP} 2**), we found that both aspects were indeed used, and even though the input aggregation modes did not excel in how they were perceived, they were still used. We thereby found that the typical plurality/majority voting schemes in polls, which is used in many live-streams (see Section 5.2), is not the one that was most often used. The group instead selected aggregators in which either an individual received full control or in which the ones with more “expertise” should have more impact. This is interesting, as it shows that individuals appear to be open to give up their influence to reach a higher goal within the game’s setting. From the answers to the game element aspects, we also learned that more means to self-administrate in such a setting are considered as valuable and that participants demanded further options to empower individuals. We also saw that the *TPP++* experience cannot be considered as fun overall. We attribute this to the fact that, compared to non-shared-control games, individuals do not have full control over the game and are highly dependent on others. In this sense the individual autonomy is limited (although the game elements added more options for individuals). Even in the situations where one player received full control, whether or not he or she remains leader depends on the group. These aspects are directly related to the thesis’ scope (**RQ3**; see Section 1.4).

While we have received insights into how viewers perceive a shared game control setting which is similar to *TPP*, we cannot reason about how effective the group in this setting was. The progress made was perceived as low by the players. From an objective standpoint, we need to consider that the chosen game setup was a role-playing game. This offers a large variety of options and goals that individual players might have (e.g., catching a new creature, leveling one up, defeating a certain NSC, exploring the area X before Y, etc.). This leads to two issues: first, there is no metric that defines exactly how to calculate the effectiveness of the group decisions here. Even though specific actions could in theory be rated (e.g., going left is faster), contextual factors and the range of potential sub-goals make a metric for assessing individual actions too complex or even impossible (e.g., in the given example, going right might earn the player an optional item that might help later in the game). Second, this range of options makes the group decisions more difficult, as individuals might pursue different objectives. Heeter et al. summarized it as “*each specific player choice is idiosyncratic to a particular moment of play in a particular game, and therefore specific choices rarely serve to characterize game play in a conceptually meaningful or even useful way. Player choices make game play interesting but they complicate play style measurement.*” ([110], p. 44). Overall, while we could investigate the *TP* setting in general with *TPP++*, the next section presents a context which allows us to reason about the quality of individual and group decisions in shared game control.

6.3 Effectiveness in a Shared Game Control Setting

To be able to reason how effective individual and group decisions are in a shared game control setting, a “simpler” context than *TPP++*, which allows evaluations of actions was needed. As games usually have multiple goals, evaluating the available game actions in terms of whether they are helpful towards these goals is not trivial (see Section 6.2.3). *Chess* software today allows for evaluating moves and *chess* has a complexity that makes playing it interesting enough to attract viewers on live-streaming platforms (e.g., on *Twitch* several *chess* channels exist that attract several hundred viewers on average⁵³). Furthermore, in “*Kasparov versus the World*”⁵⁴, over four months, people could play *chess* against the *chess* grandmaster Garry Kasparov over the Internet (but not in a live-streaming context): they could submit move suggestions and based on a plurality voting scheme, the most suggested move was carried out. According to the article, more than 50,000 people participated in this match. Altogether, we judged *chess* as a good fit to be used in a *TP* context, leading to the following goals:

- Goal_{CC} 1** *Creation of a shared game control setting to measure decision quality:* The game of *chess* needs to be adapted for the *TP* setting by accounting for the limitations in the live-streaming and shared game control context. With this game, as discussed above, the effectiveness of individual and group decisions can be measured. As *TPP++* was evaluated without considering issues of live-streaming platforms, this will also provide further insights into the overall topic of *TP*.
- Goal_{CC} 2** *Evaluation of individual and group decisions in a shared game control setting:* With the created system, we can evaluate the individual and group decisions to learn about their quality in shared game control.

6.3.1 Concept and System Design of *CrowdChess*

To account for **Goal_{CC} 1**, we developed *CrowdChess*. Similar to our considerations of *Helpstone* (see Section 5.3.2) and in contrast to *TPP++* (see Section 6.2.2), we decided to establish *CrowdChess* directly for the live-streaming context. Thus, we faced several challenges by going for an “in the wild” setting. One challenge is how to enable the audience to enter the game inputs. Even though live-streaming platforms such as *Mixer* allow channel owners to customize their channel page with HTML input elements (such as buttons) and to programmatically listen to viewer interactions, these options were limited⁵⁵. For example, it would not have been possible to provide a virtual representation of the *chess* board to allow viewers to suggest moves via direct interactions, similar to other *chess* games, while providing them with feedback at the time of the study. Although *TPP++*

⁵³ For example, <https://www.twitch.tv/chess>, (last accessed: 2018-07-07)

⁵⁴ Wikipedia: *Kasparov versus the World*, <https://goo.gl/91ubzv> (last accessed: 2018-07-07)

⁵⁵ The *Twitch* extensions (see Section 2.5.1) were not available at the time of the system creation.

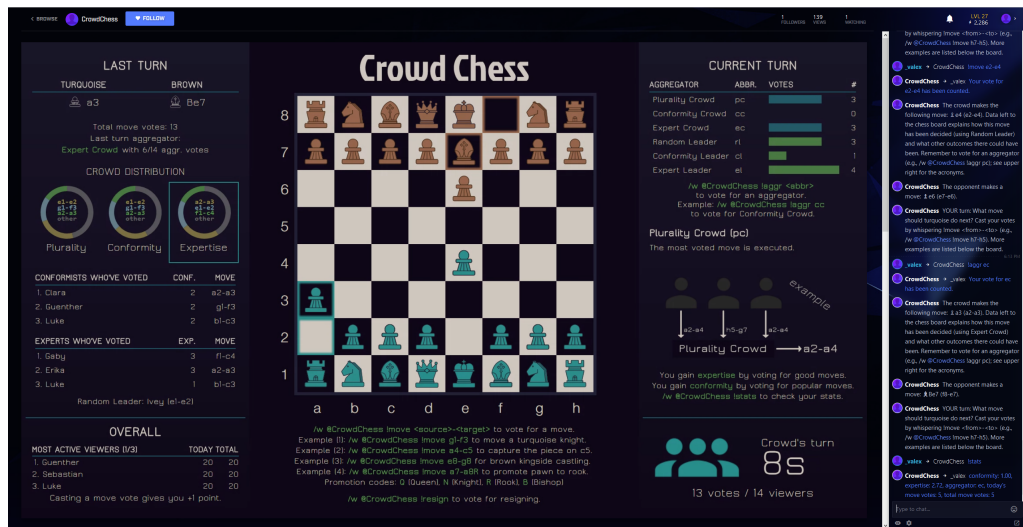


Figure 6.5: A screenshot of *CrowdChess* running on the live-streaming platform *Mixer*. Left: Our video stream. Right: The channel's chat.

and *Helpstone* worked on an external page, we decided that *CrowdChess* should be directly usable on the live-streaming platform itself, similar to the other available *TP* settings. Another challenge, especially on *Twitch* and *YouTube*, is the lag (as discussed in Section 2.5.1), making it difficult to relate chat messages to the situation seen in the stream. In settings in which audience input is interpreted as game commands, this is especially problematic and needs to be considered in the underlying game concepts. Finally, using the chat to provide information (as it works in real time, in contrast to the video stream and might thus be a reasonable alternative) to game situations and individual feedback poses the challenge that platforms have limitations on how many messages can be sent in a minute. Although white-listing options might be granted, this cannot be assumed unconditionally. Additionally, a white-listed channel might still be too limited, depending on the channel size. Exceeding the limits can, depending on the platform, mean that the messages are simply not sent to other viewers, or in the worst case, that the sending account is banned.

With respect to these challenges, we created *CrowdChess* as a test-bed that can be used directly on today's common live-streaming platforms (see Figure 6.5) in an “in the wild” fashion. It consists of four components: the *chess* engine, the aggregators, the live-streaming integration and the user interface shown in the stream. *CrowdChess* is designed for audience vs. Artificial Intelligence (AI), to investigate **Goal_{CC} 2**, but the system can easily be adapted to allow for streamer vs. audience or audience vs. audience matches, which might also reveal interesting results. As *chess* is already round-based and partitioned into turns and does not require fast interactions, this helps to lower the impact of the lag issue mentioned before. The idea of *CrowdChess* is that every user can suggest a move per turn and through input aggregation (similar to *TPP++*) and coordination amongst the players, one of the suggestions is carried out as the group result.

Chess engine and AI: We use *Stockfish*⁵⁶, an open-source *chess* engine, to evaluate the individual and aggregator-derived move suggestions. Besides using these results as data points later, we use them to inform the expertise-based aggregators (see below). The engine is too strong to be a motivating challenge in typical live-streams. Therefore, we use some of the engine-offered options to limit it (such as the maximal search depth for possible moves): *CrowdChess* offers 16 AI levels; the level increases (decreases) by one when the audience wins (loses). The lower levels can be easily beaten even by *chess* beginners (as tested informally).

Aggregators: *CrowdChess* allows the audience to alter the way individual move suggestions are aggregated by voting for one of six aggregator schemes, which are based on our *TPP++* setup. We again distinguish between two concepts: those that aggregate the input (named “crowd”) and those that select an individual based on certain attributes and use the related move suggestion (named “leader”).

- **Plurality Crowd (PC):** A plurality vote, i.e., the move suggested by the most viewers is executed.
- **Expert Crowd (EC):** A weighted plurality vote with weights based on individual expertise levels. How the expertise is calculated specifically is not disclosed to the viewers; they can only see the rating, which is updated for every move suggestion given. Every viewer starts with an expertise of 1. A given move suggestion is evaluated by the *chess* engine in relation to the current board situation. This increases (decreases) the viewer’s expertise by up to 2 (-2) points depending on how good (bad) the suggestion was. For this, it is considered whether the suggestion leads to the opponent’s mate, to one’s own mate (provided other moves would have been possible), whether the viewer has missed mate in 1/2/3 turn(s) and finally, how the suggestion score relates to the score of the best possible move in the given board situation. An audience having the goal to beat the AI should empower experts, as they more likely provide good move suggestions. As the Expertise Weighted Vote aggregator had the highest uptime in *TPP++*, it will be interesting to see whether this is replicable here.
- **Conformity Crowd (CC):** This aggregator is also a weighted plurality vote where the weights are based on the individual conformity levels. Move suggestions given by viewers are continuously compared to suggestions provided by other viewers in the same turn. Suggesting moves that were also suggested by the majority increases a viewer’s conformity by up to 2 (starting at 1), depending on the uniformity among move suggestions. The level is decreased (by up to -2) when suggestions are made which only small portions of (or no one else in) the audience also made. As soon as Expertise Weighted Vote became available in *TPP++*, the conformity-based Crowd Weighted Vote aggregator had nearly no uptime anymore. Thus, it will be interesting to see whether the aggregator will be used at all here.

⁵⁶<https://stockfishchess.org> (last accessed: 2018-07-07)

- **Random Leader (RL):** The random leader aggregator randomly selects one of the viewers' move suggestions. Such an aggregator relates to the "*anarchy mode*" in the original *TPP* (or *Mob* in *TPP++*), in which all commands were executed and thus all viewers could contribute something. Using the concept of the "*anarchy mode*" is possible, but seems unreasonable for a *chess* setting, as the chance of doing something "worse" and not being able to reverse this decision through other viewer commands, in contrast to the game *Pokémon Red*, is much higher. RL in contrast ensures that in theory all viewers can participate in the game independent of their conformity/expertise level, but at the same time also allows the matches to still be playable.
- **Expert Leader (EL):** In contrast to RL, EL uses the move suggestion of the viewer with the highest expertise. In *TPP++* the audience empowered individuals in difficult situations (but no direct counterpart to EL was available in *TPP++*). In *chess*, it seems reasonable to offer an aggregator that provides the best player with full control. Similar to EC, empowering the best user seems reasonable, when the goal is to beat the AI.
- **Conformity Leader (CL):** In contrast to EL, here the conformity level is the selection criterion (i.e., the move suggestion of the active player with the highest value in conformity is the one used). In *TPP++* this kind of aggregator was also available and was used in difficult situations to provide an individual with complete control. It is interesting to see whether this aggregator is used when EL is offered in parallel.

The selection of the active aggregator is an ongoing plurality vote. Viewers can switch their aggregator vote anytime and the one with the most votes becomes active at the end of a turn. The time left in a turn is always shown to the viewers. In the case of a draw, one of the aggregators involved in the draw is randomly selected. If the active aggregator provides the same value for multiple moves, one of these is also randomly selected.

Live-streaming integration: *CrowdChess* uses a chat bot to inform viewers about their interaction options via the channel's chat. These messages also provide insights into the current game state (e.g., whose turn it is and which piece was moved last). Especially for the lag issue on the live-streaming platforms, this allows the viewers to think about the next turn, even if the video stream does not yet show the corresponding state visually. Example messages are:

Your turn: What move should turquoise do next? Cast your vote by whispering !move <from>-<to> (e.g., /whisper @CrowdChess !move h7-h5).

The crowd makes the following move: pawn h4 (h2-h4). Data to the left of the chess board explains how this move has been decided (using Random Leader) and what other outcomes there could have been. Remember to vote for an aggregator (e.g., /whisper @CrowdChess !aggr pc); see upper right for the acronyms.

We require viewers to interact with *CrowdChess* via whispers. We did not want viewers – for this consideration of the effectiveness of individuals and groups – to influence others merely by disclosing what they want to do, simply by displaying their vote command publicly. Such social effects were already reported by other research [179, 184, 188] and should be reduced with whispering. These whisper messages are also the primary feedback channel of *CrowdChess*. We use this channel to acknowledge that a viewer has correctly entered a move suggestion, or to inform him or her that a move is not possible in the current board situation or that an ill-formed command was entered. We implemented a queuing system to cope with the situation of having no white-listing and to avoid exceeding the limits the live-streaming platforms allow. If viewers enter commands into the chat instead of whispering, the chat bot removes the entered command from the chat and sends a message informing all users that they need to whisper. Valid move suggestions follow the form *!move* <from field>-<to field>. This notation is used for all kinds of move suggestions in *CrowdChess* (e.g., capturing other pieces and castling) to simplify the interaction. A user is able to enter multiple ones per turn, but only the last one is considered at the end. By entering *!aggr* <aggregator name> players can switch their aggregator vote. *!stats* provides players with their current statistics (i.e., their conformity/expertise level, which aggregator they have currently voted for and the numbers of votes entered today and overall). Entering *!resign* is treated as a special move suggestion. If the active aggregator selects this suggestion, the match ends.

We use the channel description to give an overview on the command options and the aggregators. Together with the streamed interface of *CrowdChess* (see below) and the chat bot's proactive and reactive messages, viewers have three sources to learn how to use *CrowdChess*. Finally, the game is designed to run continuously: as soon as a match is over (or no viewer is available anymore), the game restarts with an adapted AI level (as explained before) and is ready for a new match.

User interface: Figure 6.6 shows the user interface consisting of historical information (left), the *chess* board (center) and the current information (right).

Historical information: Here, data on the last turn and the overall game is shown. At the top the last move of the audience and the AI is displayed. Below, the active aggregator is shown and how many viewers wanted it in relation to all given aggregator votes. This is followed by a visualization of the moves the aggregators would have selected (the last active aggregator is additionally marked with a cyan box). This should help viewers to make informed future aggregator vote decisions. Pie charts are used for the crowd aggregators. Each chart displays the top three moves the aggregator outputs in the middle, and their relative distribution in the circle. For EL and CL, tables show the top three most conforming/experienced viewers with their suggested move in the last turn. Which viewer would have been (or has been) selected randomly by RL and his or her move suggestion is also denoted here. Finally, the lower area switches every 20 seconds between showing the top nine active viewers (in groups of three) and how many matches the audience and AI has won today/overall.

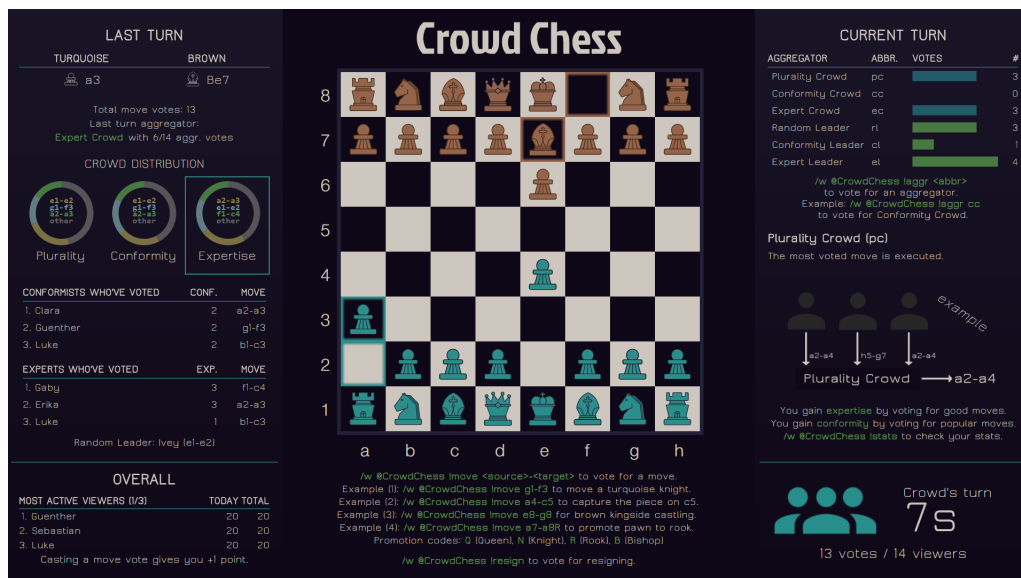


Figure 6.6: The user interface of *CrowdChess* consists of last turn information (left), the *chess* board (middle) and the current turn information (right).

Chess board: In the center the *chess* board is shown together with the corresponding column letters and row numbers as these need to be entered by the viewers in their move suggestions. On the board, the last moves made are highlighted with rectangles in the color of the audience and AI, to also help later-joining viewers. Below the board, examples on how to enter move suggestions are given. It is shown how to enter a normal move, how to capture pieces, how to enter special moves (castling/promotion) and how to resign the game.

Current information: Information relevant for the current turn is shown on the right side. The upper part displays the aggregators' distribution by showing the different aggregator names (and their abbreviations) and how many viewers have voted for every aggregator. This is followed by a short explanation on how viewers can switch their aggregator vote and a large area providing an example for every aggregator. Every 20 seconds a different aggregator is explained here (see Figure 6.7). With this element, viewers do not need to read the channel description to understand how aggregators work. We also give information on what expertise and conformity mean in our context, and provide the information that one's own expertise/conformity value can be queried with the *!stats* command. Below this area, it is shown whether the audience or the AI needs to make a move. When it is the audience's turn, the remaining turn time is also displayed. To further mitigate the lag issue, we distinguish between lag and turn time. The lag time denotes a time before the turn timer starts. The lag and turn time can be adjusted, but are fixed at runtime (i.e., the audience cannot vote to set up the time by themselves). Below the time it is indicated how many viewers are present and how many have entered a move suggestion currently.

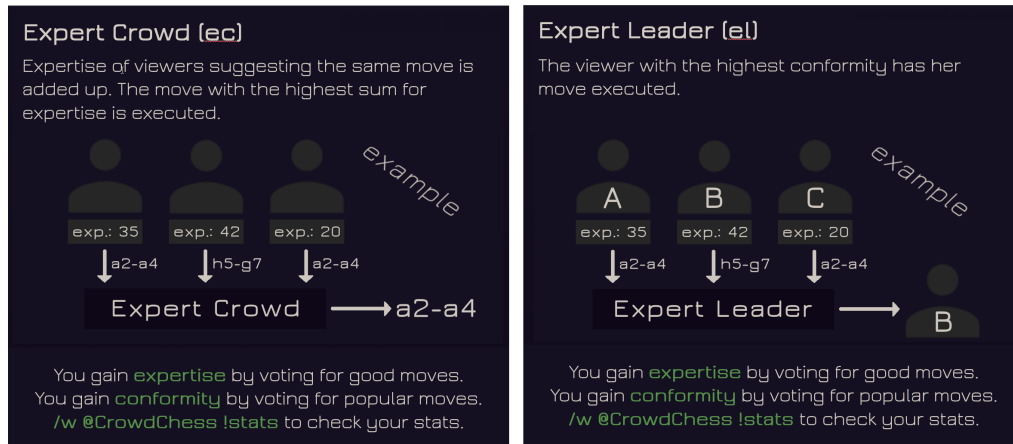


Figure 6.7: The explanation shown for the EC and EL aggregators.

6.3.2 User Study with *CrowdChess*

We conducted an exploratory user study, where we were interested in how *CrowdChess* is perceived (especially in relation to the findings of *TPP++* reported in Section 6.2.2), how the players use the aggregators and how effective their game decisions are in a shared game control setting (related to **Goal_{CC} 2**).

Method

Although *CrowdChess* is implemented to work with arbitrary live-streaming platforms, we decided to use the platform *Mixer* for this study because the lag is reported to be below one second there (see Section 2.5.1). Thus, viewers would nearly instantly see moves on the streamed video and are not restricted to the chat-bot until the lag time is over. Simply launching a channel on a live-streaming platform was shown to be problematic in several related works that crawled platform data (e.g., [139, 334]): most of the channels on live-streaming platforms do not attract many viewers in parallel, or any viewers at all. Therefore, we advertised an event in which “*You should play chess as part of a group on a live-streaming platform to beat an AI*” over *Facebook*, chess communities and student mailing lists (consisting of computer science, media informatics and psychology students). As prerequisites, we mentioned that players should know the rules of *chess* (but that skill level is irrelevant) and that a *Mixer* account is necessary. During runtime it was also possible that users joined who were not explicitly recruited in this way. In general, this kind of advertisement is similar to the usual live-streaming case in which streamers also, for example, announce their streaming times and content via social media platforms.

The event ran for 45 minutes. Before it began, we streamed a slide with information on the event (such as that the channel description provided all necessary commands). Here, we also stated that all communication should be done over the

channel's chat and not over tools such as *Skype*. The turn time was restricted to 60 seconds (and one second lag time), leading to 61 seconds between the opponent move being shown in the chat and the next turn of the opponent. The time restriction was to allow a faster game play and ensure that more moves could be carried out during the event. The AI was initially set up with a depth of five (easy) and was then set to ten (rather challenging) after the first match, to let the audience play against an easy and a harder opponent. After the 45 minutes, we provided a link to a questionnaire (which was available in English and German) for the participants via the channel's chat. It assessed demographics, a self-assessment of one's own skill level, the perception of *CrowdChess* based on statements to be answered on 4-point scales (with the labels *disagree*, *rather disagree*, *rather agree*, *agree*) and some optional questions that allowed for free-text answers. We logged the chat messages, recorded the live-stream and persisted all interactions with our system for analysis later on.

Results

We clustered the results into general usage statistics, qualitative feedback, move suggestion quality and aggregator uses.

General usage statistics: 18 registered users visited our channel during the experiment: 13 wrote at least one message (independent of whether it was a chat message or command) and three users remained as spectators for an average of two turns (without doing any interactions before leaving the channel); the other two users visited and left immediately. One user only entered chat messages (this user joined a few seconds before we closed the event). The remaining twelve users all entered at least one move or aggregator change command, i.e., they participated directly in the game (and will be called "players" subsequently). Nine of them used the chat to communicate (and wrote a total of 153 chat messages). The players entered 214 move suggestions ($M=17.8$, $SD=8.1$, $Mdn=21$), 24 aggregator commands ($M=3.4$, $SD=2.3$, $Mdn=2$; done by seven players) and eight stats requests ($M=1.6$, $SD=.9$, $Mdn=1$; done by five players).

Two matches were finished: the first one against AI level five was won by the audience, the match against AI level ten by the AI. Ten players played both matches; two players entered player commands in the second match only. Overall, the audience had 31 turns. In match one (13 turns) the average number of move suggestions per turn was 7 (min=4, max=9); in the second match (18 turns) 6.8 (min=3, max=11). Table 6.2 shows the number of actions per player, their average expertise/conformity value and percentages on how active they were in relation to the number of turns they were on the channel. 75% of the players participated in more than 50% of their witnessed turns.

Twelve participants (seven male, three female, two no answer; age: <19: 2, 19–25: 5, 26–32: 2, 33–39: 2, no answer: 1) finished the questionnaire and provided us with their *Mixer* account name. We could thereby confirm that all active players

ID	Number of actions				Avg. EX	Avg. CO	Seen turns	% of turns in which the user did a player action	% of witnessed turns in which the user did any action	Skill self-assess-ment
	M	A	S	C				<i>!move/!aggr</i>	command/chatted	
01	27	6	0	8	17	13	31	87%	97%	SE
02	21	7	0	2	4	5	31	74%	77%	IN
03	26	0	3	18	10	11	31	84%	87%	IN
04	18	1	0	11	13	10	31	61%	65%	SE
05	6	2	0	10	2	3	31	26%	39%	IN
06	15	2	2	13	6	5	31	52%	71%	SE
07	22	0	0	0	7	11	31	71%	71%	IN
08	23	4	1	28	12	9	31	81%	97%	SE
09	24	2	1	61	9	7	31	81%	97%	IN
10	21	0	0	0	6	9	31	68%	68%	IN
11	8	0	0	2	3	2	23	35%	39%	IN
12	3	0	1	0	2	2	11	27%	27%	SE

Table 6.2: Overview on players in our *CrowdChess* study. Abbreviations: M=Move, A=Aggregator, S=Stats, C=Chat, EX=Expertise, CO=Conformity, SE=Some experience, IN=Inexperienced. As players were able to join the game later, we added indications on how active they were in relation to their witnessed turns.

did the questionnaire. All but one (Scottish) reported to be German. The Scottish player joined the event by coincidence, while the remaining players heard about it via our advertisement. We let the participants rate their own skill level (see also Table 6.2) by providing them with statements that indicated different skill levels sorted from lowest to highest. They either selected “*I know the chess rules, but I have not played chess much so far*” (which we denoted with “inexperienced” (IN) in the table) or “*I know the chess rules and I think that I can win against other casual chess players*” (“some experience” (SE)). The next statement “*I know the chess rules and I think that I can win against players that play chess regularly but are not chess club players*” was not selected, indicating that the players were not skilled chess players. No one reported playing chess regularly. Only four participants reported consuming live-streams, and three had participated in the original *TPP*.

Qualitative feedback: Seven participants stated they liked *CrowdChess* (three times agree, four times rather agree), with an average of 2.8 on the 4-point scale ($SD=1$, $Mdn=3$). Seven participants reported having fun (only asked on a single scale) playing it ($M=2.6$, $SD=1$, $Mdn=3$) and four participants stated that it was more fun than playing “normal” chess ($M=2.2$, $SD=.9$, $Mdn=2$). The idea of playing chess as a group was liked by half the participants ($M=2.5$, $SD=.8$, $Mdn=2.5$), and five would continue to play *CrowdChess*. Potential aspects that might have impacted the perception of *CrowdChess* were:

- **Time:** No one reported being familiar with playing *chess* under time pressure. Eleven participants rather disagreed or disagreed with the statement that the turn time was too long ($M=1.3$, $SD=.6$, $Mdn=1$) and ten disagreed with the statement that they had enough time to play and consider the historical information ($M=1.3$, $SD=.8$, $Mdn=1$). In the free texts, six reported that they wanted to participate in every turn and thus had insufficient time to consider all parts of the interface. This study tried to mimic the live-streaming setting in which visitors also simply join the channel and are confronted with it directly. Therefore, we did not explain the user interface beforehand. As the study time limit of 45 minutes was communicated clearly, participants might have thought that they needed to provide a move in every turn (which was not necessary) or found it more appealing to enter move suggestions. Concerning time, two participants suggested in the free-text answers to either add an automatic turn time adaptation (by considering how much discussion happens in the chat) or audience-based options to adjust the time. They also reported that they wanted an option to skip the remaining time, if the audience has already decided. These suggestions indicate that at least some participants want to have more influence in these group settings, i.e., influence that provides them with more autonomy to change system aspects, fitting the goal of this thesis.
- **User interface:** Seven participants at least rather agreed (i.e., selected either rather agree or agree) to the statement that they like the graphical appearance of *CrowdChess* ($M=2.4$, $SD=1$, $Mdn=3$). In addition, nine participants found the chat bot messages useful ($M=2.9$, $SD=.9$, $Mdn=3$) and nine liked to see the current aggregator distribution ($M=3$, $SD=1.1$, $Mdn=3$). Five times it was mentioned in the free-text answers that the UI is too overloaded and three times it was reported that the user's monitor (and thus the video stream window) was too small to see all elements properly. Seven at least rather agreed to the statement that they understood the aggregator explanations in the stream ($M=2.7$, $SD=1.2$, $Mdn=3$) and nine at least rather agreed to having understood the aggregator distributions ($M=3.2$, $SD=1$, $Mdn=3.5$), but only four at least rather agreed that they understood all components of the history ($M=2.2$, $SD=1.1$, $Mdn=2$).
- **Text-based interaction:** Seven participants at least rather agreed to the statement that the interaction via chat messages was acceptable ($M=2.4$, $SD=1$, $Mdn=3$). Four participants explicitly stated that they would rather interact directly with the *chess* board on the stream. This is in line with the findings we gained in *Helpstone* (see Section 5.3.2), i.e., that the direct interaction was perceived as beneficial. Nine participants at least rather agreed that they always knew how to enter a move suggestion ($M=3$, $SD=1.1$, $Mdn=3$). Six syntactically incorrect and 41 invalid (with respect to the *chess* rules) move commands were entered, but considering who entered those, no relation was found. Even though the chat bot provided an error message via whisper, this might have been overlooked.

- **Group feeling:** Two participants reported that they were frustrated because moves were executed that they assessed as worse than their suggestions. While five at least rather agreed to the statement that they felt part of the group ($M=2.3$, $SD=1.1$, $Mdn=2$), six reported that they thought that their move suggestions mattered ($M=2.3$, $SD=1.2$, $Mdn=2.5$). Two participants suggested replacing the historical information with the current move suggestions all players have provided in the turn to come to a better decision.

Eight participants were interested in the expertise ($M=2.6$, $SD=1.1$, $Mdn=3$) but only two in the conformity scores ($M=1.6$, $SD=.8$, $Mdn=1$). None of the participants reported using the stats request command frequently to check his or her expertise ($M=1.1$, $SD=.3$, $Mdn=1$) or conformity ($M=1.1$, $SD=.3$, $Mdn=1$) (also in line with the actual frequency of the `!stats` command) and only five (one) reported checking the table in the history to see the expertise (conformity) distribution. As an optional question we asked whether the expertise/conformity values derived by *CrowdChess* seemed plausible; while the three participants that answered this at least rather agreed for conformity ($M=3.3$, $SD=.6$, $Mdn=3$), three (of four) participants that answered this for the expertise value at least rather agreed ($M=2.8$, $SD=.5$, $Mdn=3$).

Move suggestion quality: We first analyzed the 214 move suggestions for how different they were for every turn. We counted how many groups of the same move suggestions in a turn were provided and divided this by the number of all suggestions in that turn. Averaged over all turns, a number near 0 would indicate that in most turns the audience would have provided only uniform suggestions; a number near 1 would indicate that the audience only provided different move suggestions in most turns. Both would render most of the aggregators useless, but this was not an issue in our case ($M=.5$, $SD=.2$, $Mdn=.4$). Using the *chess* engine, we analyzed all 214 move suggestions. Only 40% of the suggested moves would have changed the board situation in favor of the crowd. This is in line with Table 6.2 and the skill self-assessment. For every board position (31) we aggregated the given move suggestions by using all six aggregators regardless of which one was actually active in the turn. For the random leader aggregator, we used the suggestion of the player that was randomly selected in this turn; for the expert/conformity leaders we utilized the values that the players had at the point in time when the suggestion was made. Based on this, in 25 of 31 turns (81%) at least one aggregator provided the best result amongst the user suggestions (i.e., it selected the move suggestion that received the highest engine score). Table 6.3 shows the performances of the aggregators. The table represents the case without filtering situations in which the worst was the same as the best move suggestion (e.g., all move suggestions were equivalent). Column 2 of this table shows that the offered aggregators are not perfect. PC, for example, only selects the best suggestion in 64.5% of the turns. This is not surprising as this aggregator was only able to select the move when the majority of the players suggested it. Even the expert aggregators did not clearly excel here, indicating that no player (especially the ones that had a slightly higher expertise value)

Agg.	% best move suggestion selected	Avg. % of move suggestions better than move selected by aggregator (across all turns)	% worst move suggestion selected
PC	64.5%	13.6%	29%
CC	71%	13.1%	19.4%
EC	71%	11.7%	25.8%
RL	51.6%	24.8%	32.3%
CL	64.5%	17.8%	29%
EL	71%	15%	29%

Table 6.3: Aggregator performances based on all board positions (31) and all move suggestions (214).

consistently provided better moves than other players. In column 3 we present an average value: for every turn, we checked how many move suggestions were provided that were better than the selected one. Here, the “crowd” aggregators seem to be better than the “leader” aggregators. This is explainable with the concept of the *wisdom of crowds* [289], i.e., that a group of people comes to a better decision compared to an individual decision. Considering the absence of skilled *chess* players, this seems reasonable. Column 4 shows how often an aggregator selected the worst suggestion, with RL being (as expected) worst.

Aggregator uses: Table 6.3 shows only a generalized view that does not consider when an aggregator was actually activated by the audience. For example, if an aggregator outputs the best move amongst all available suggestions only 40% of the time, it does not necessarily mean that this aggregator is bad. If the audience selected this aggregator only in situations in which it outputs the best move of the available user move suggestions, this would lead to two conclusions: first, it would show that the audience is able to self-administrate itself (by knowing which aggregator is currently good) and second, that even though the overall performance of the aggregator is suboptimal, in terms of how it is used, it is optimal. We found that in 20 of 31 turns (65%), the audience indeed activated an aggregator that selects the best move amongst the user suggestions. Considering that in only 25 turns the aggregators could have provided the best outcome, this is an encouraging result. Table 6.4 shows how often an aggregator was active and its performance. These numbers also represent the case without filtering situations in which the worst and best move suggestion were the same.

The following results could be derived from Table 6.4: first, the most active aggregator was EC. This is in line with our expectation (see aggregator explanation in Section 6.3.1). Interestingly, EL was never activated. It seems that the audience still wanted to empower their members to contribute something instead of simply providing one expert with the option to play alone. Another explanation might be that they were aware that no single player alone excels. Second, activating EC was a reasonable decision, as it selected the best suggestion in 13 of the 17 turns it was active. We analyzed what happened in the other four turns: two times at least one other aggregator would have provided the best result, namely PC for

Agg.	#Turns active (# of last active turn)	Draw wins	% best move suggestion selected	Avg. % of move suggestions better than move selected by aggregator (across all active turns)	% worst move suggestion selected
PC	7 (#16)	5	57.1%	13%	14.3%
CC	0 (-)	0	-	-	-
EC	17 (#31)	2	76.5%	14.3%	35.3%
RL	5 (#19)	5	40%	31.6%	60%
CL	2 (#2)	2	100%	0%	0%
EL	0 (-)	0	-	-	-

Table 6.4: Aggregator performances while active in a turn.

one turn and RL for the other. As choosing RL could not be accounted for as a deliberate “better” choice, this led to only one turn in which another aggregator would have provided the best result. By additionally considering how often any aggregator would have provided a better result (and not necessarily the best), one instance of RL and one instance of PC/CC/CL would have. Second, compared to Table 6.3, EC and CL performed (slightly) better, i.e., they seemed to be activated at the right time. Third, we had many turns (14/31) in which two or more aggregators had the same number of upvotes (draws). PC, RL and CL had most (or all) of their uptime only because they were randomly selected in such situations. This shows that the crowd was not uniform. EC instead was preferred clearly as it was active in 17 turns (with only two draw wins). EC also dominates in the second match, as PC for example was never activated after the 16th turn. This dominance was also visible in the vote distributions, as four participants wanted EC, while only two wanted other aggregators towards the end of the study (4:1:1). Fourth, the conformity-based aggregators were not really used (even though CC is about as good as EC, as shown in Table 6.3). Participants also stated that they were not interested in the conformity. Thus, they either did not understand the functionality of the value itself, did not understand the aggregator, or they found the expertise-based aggregators more appealing. As seven participants stated they did not feel part of the group, they might have thought that CC and CL do not provide coherent results, as they are “outside” the group. We also checked for every aggregator how many players voted for it at least once: PC (3x), CC (1x), EC (7x), RL (3x), CL (1x), EL (0x). This again shows that CC, CL and EL were not appealing for the players and that EC dominated.

Overall, these results should be seen in the context that only seven players did aggregator switches at all. As we had a question on infrequent aggregator switches in the questionnaire (and allowed multiple selections) we could investigate this further: considering the players that never used the aggregator change command, the players with IDs⁵⁷ 10 and 12 reported that they did not understand the aggregators, players 03 and 11 that they had too little time and players 03 and 07 that they thought switching the aggregator would not have led to better moves and

⁵⁷ Indicates the ID in Table 6.2.

that they were satisfied with the current aggregator. Considering the players that voted for aggregators, 01, 04 and 05 reported that they had too little time, 02 and 06 that switching would not have led to better outcomes and 01, 04, 08 and 09 that they were satisfied with the current aggregator.

Participants were also asked why they did change the aggregator: 06 stated that RL was more interesting as he had not expected to make an impact in EL and CL. 08 stated that he was interested in PC and EC and that the game was too fast to switch the aggregator more often. 01 reported that he switched *“when moves were executed that I thought were bad and a switch would favor mine”*; 09 stated *“partially, I have selected expertise, because I’m not so good myself”*. 02 answered that she wanted to prevent a move being executed randomly. This shows that the motives for why changes happen are different. None of the participants thought that *CrowdChess* needed further aggregators, and four thought that there were too many.

We also considered the chat to learn which social interactions towards decision-finding happened that might be a further explanation for why the aggregators were not used more often. 15 turns were discussed in terms of which move should be done next (the chat was used in both matches for these discussions): either by a user giving a concrete suggestion (e.g., 06: *“f8 - b5”* or 01: *“I recommend to move the queen”*), a player starting a discussion (e.g., 03: *“g8-e7?”*), a player asking for help (e.g., 09: *“suggestions from pros, please”*) or discussing more general plans (e.g., 06: *“cover the e5”* or 09: *“killllllit - with the pawn?????”*). Seven players participated in these discussions. One time, a player asked which aggregator should be activated and two other players responded with their preferences. Similar to *TPP* [187], even in this small user base, we found trolling tendencies by one user (e.g., 06: *“we should always do the same like the AI”* or 06: *“randomize it”*).

Discussion

In this study, we only tested *CrowdChess* with a small number of participants in an exploratory fashion. Our results should thus not be overestimated, but they already give valuable insights on how small groups of viewers interact in such a scenario: first of all, we found that the aggregators were not used by all players. As we could show, EC performed well in this setting, and the audience mainly activated this aggregator. Viewers might have voted for EC because they liked its move selection, they approved our expertise metric and/or they liked the aggregator strategy. As no other aggregator significantly outperforms EC, we cannot decide which hypothesis should be accepted, but this can be further explored with *CrowdChess*: for example, by adding a “fake aggregator” that is purely based on the *chess* engine and thus can (more easily) outdo the other aggregators. In contrast, PC – often used in the live-streaming context – was not often favored (otherwise it would not have come to so many draws) and had no uptime later on. The conformity-based aggregators were not interesting for the audience, even though the performance of CC was slightly better than EC’s. As many participants reported not feeling part of the group, this might

be an explanation of why the audience had no trust in an aggregator that uses one's similarity in relation to the group. In *TPP++* we saw that the conformity-based group aggregator was not interesting anymore after an expertise-based aggregator became available. Both show that conformity-based aggregators are less interesting as soon as other aggregators become available. Interestingly, while in *TPP++* the expertise values were viewer-based (viewers could upvote others), *CrowdChess* calculates these objectively. In both cases, aggregators based on these values were perceived positively and were used often. Furthermore, EL was never activated, indicating that either the audience has more trust in the *wisdom of crowds* [289], or that they simply did not want to enable one player to decide the moves alone. Both are also supported by considering that the "leader" aggregators had less uptime in comparison to the "crowd" aggregators. This is a difference from *TPP++*. One explanation for this might be the fact that in *TPP++* difficult situations become obvious (and in consequence leaders were elected), while in *CrowdChess* identifying difficult situations is harder and needs a certain experience with the game, which our sample might not have had.

From the qualitative answers, we learned that there were different reasons why participants did not vote for aggregators (more often): besides too little time, they often reported that they were already satisfied with the active aggregator or thought that other (inactive) aggregators would not perform better. This seems to be a reasonable explanation and hints that this needs to be considered in such shared game control settings. Not participating in something can also mean that the users agree with what is happening, and today, when polls are used in streams, the number of non-voting viewers is not considered. We also learned that there are different motivations for why aggregators should be changed. Even though EC was selected, other tendencies were also revealed that are not primarily helpful for the goal of winning the game. Considering the original *TPP*, this might explain why so much chaos happened there and led to the introduction of the plurality-based aggregator ("*democracy mode*"). Although participants reported problems of time, it seems that some still had enough time to discuss moves in the chat, instead of relying on the aggregators alone. We hypothesize that this has less impact in larger channels, as work already exists that shows that the chat in larger channels is hard to maintain [108] or that the chat dynamics change [79] (see also Section 2.5.1).

Our study had limitations: first, the small number of participants, which seems acceptable for a first exploration with *CrowdChess*. We expect that further studies with this test-bed and a larger sample might change the dynamics that happen in the chat. Also, the selection of aggregators might be different when more users have trolling tendencies or want to follow their own agenda. Here, it will be interesting to see whether participants who have not voted so far start to vote to reduce the impact of such players. Second, restricting the study time to only 45 minutes meant participants were only able to play two matches and were keen on participating in both, instead of interacting with the other *CrowdChess* features. Third, regarding the selection of our sample, only four participants had

live-streaming experiences, and no one was a regular *chess* player. Conducting a study with *chess* players only, or players that are regular consumers of live-streams, might reveal different results in relation to the aggregators (for the *chess* group) or towards shared game control settings (for the stream viewers). Fourth, we investigated the board game *chess* instead of digital games. Even though we explicitly wanted to use a less complex game in terms of its game actions to evaluate aggregator decisions, it is currently unclear how these findings map to other (video) games.

6.3.3 Contribution to the Thesis' Questions

Overall, the study results show (in relation to **Goal_{CC} 2**) that the sample were able to self-manage themselves, as those who used the aggregators selected the most valuable one and did not empower individuals alone, indicating that the group interaction was still a reasonable goal. Overall, we conclude that shared game control settings work but that they need to provide more features for self-management than simply considering all individual commands. Based on the conducted study, we found that the group of players can indeed effectively play the game; following the *wisdom of crowds* [289] idea, they are more effectively than an individual alone, at least in our study setting. Considering the thesis scope, with *CrowdChess* we presented a system that allows shared game control and could show that a group of people that is tightly coupled can still effectively play together and self-orchestrate, further adding to **RQ3** (see Section 1.4).

As *CrowdChess* was developed as a test-bed for aggregators and mechanics in *TP*-like settings, in which the effectiveness of group decisions can be measured quantitatively (**Goal_{CC} 1**), further aspects can now be investigated. For the aggregators, it would be interesting to see whether the usage patterns remain stable across samples (size and composition). The exact rationale for how we derived the conformity and expertise values was not disclosed. In upcoming studies this could be explained in detail, and different approaches on how these values are updated could be compared in respect to the viewers' perceptions. At the same time, effects on the viewers' perceived autonomy and fairness could also be considered. Another valuable direction could be to analyze the social dynamics that happen in the chat in relation to the audience size and how this affects the move suggestions and aggregator selections. This could be contrasted with a study in which an aggregator is dictated by the system (and not by the viewers) for every turn.

6.4 Summary

In this chapter, we considered shared game control settings in live-streams. With *TPP++*, we investigate how such a setting is perceived, but also whether additional means to self-administrate are beneficial. The latter had a significant impact

on the perception of shared game control settings and provided individuals with more options. Towards this thesis' scope, it again shows that empowering individuals is beneficial and has an impact on the experience, adding to **RQ3** (see Section 1.4). With *CrowdChess*, on the other hand, we presented a system that exemplifies how to develop a shared game control setup respecting the requirements and issues for direct usage on live-streaming platforms. It also allowed us to study whether the group decisions in such a setting are reasonable and effective from an objective viewpoint. We found that this was the case. With both studies, we learned that a plurality voting scheme was not particularly relevant in the presence of other aggregation mechanisms. Instead, those that empower experts in the group were more important to the users. In addition, in *TPP++*, we found that individuals also received full control to overcome difficult game situations, something we have not seen in *CrowdChess*. In contrast to Chapter 5, where we considered the usual live-streaming case in which a streamer is present and is able to orchestrate the audience, in this chapter, we have considered a new experience for live-streaming situations in which a streamer is absent. We will conclude this chapter by summarizing how both forms relate to each other:

Moderated influence: In the conducted studies we saw that an individual contribution is not necessarily integrated in the stream. Thus, the individual alone has no direct, unfiltered influence option. This leads to viewer actions that are simply “thrown away”. In the usual live-streaming situation the moderating factor is the streamer (or associates); in the shared game control settings such a moderation can be seen as due to the aggregation system. We presented techniques with which individuals can (to a certain extent) increase the chance of their contribution being used: the studies in Chapter 5 showed that by providing high quality material (for example a good story element) or good contributions in *Helpstone*, these might be more likely to be selected (because of upvoting or similar aspects). In the shared game control settings, changing the aggregation system towards one that gives this viewer more influence, or buying elements from the shop in *TPP++*, helps to make an individual contribution matter. It seems problematic to let every viewer contribute in an unfiltered way. In the original *TPP* this was the initial idea, but was shown to be problematic in terms of progress, especially in difficult areas. Nonetheless, we were able to show that more aggregation mechanisms, and the autonomy to change which one should be active, are used by the audience. This, on the other hand, is something that also should be considered for interactivity in usual live-streams, as here polls are popular, but always, to our knowledge, are based on plurality votes only.

Emergence of new experiences: Considering the *Self-Determination Theory* (see Section 2.2.2), the conducted studies in both chapters provided insights in streams that give the audience more autonomy. The streams in Chapter 5 provided more autonomy by allowing viewers to shape the experience (see Section 5.2.1), or allowed them to give hints to the streamer in an easier fashion than the chat (see Section 5.3). All studies in this chapter empowered individuals to participate in shared game control settings. Having more autonomy and options in streams

seems interesting even for passive viewers, as seen in the online study (see Section 5.2.2) and with *Helpstone* (see Section 5.3.2). Merging the results, we could easily imagine new interaction forms that vary with the degree of autonomy and thus shape new experiences: The streamers could keep some form of meta-control and, for example, decide during the stream how the community is allowed to interact. Sometimes the audience could receive full control similar to the considered shared game control settings in this chapter (and the streamer would only comment), or could simply choose an aggregator for upcoming polls. We saw with *TPP++* that there are occasions in which the audience let individuals decide on the course of action; in such new experiences this individual could be the streamer. *Twitch* tries to encourage these new experiences by their “*Stream First*” approach⁵⁸. In general, what this also shows is that research on different types of streams can also shape other types. Moreover, this kind of research helps to better understand the need for (viewer) autonomy in live-streaming.

Freedom of choice: We learned that not every viewer wants to exert influence. We saw in Section 5.2.2 that many viewers are passive (and still appreciate the interactivity, even though they would not use it) and found indications that features are perceived differently depending on the motivations a viewer has. In this chapter we also saw differences in how viewers behave, i.e., not everyone participated equally (see for example *TPP++*). This is also a core aspect of this thesis: in giving users influence, one should always consider that users may not want to utilize the range of options provided. Similarly as we have done in Chapter 4 with “bottom-up” gamification, we reason that this should be accounted for in the experiences. Exerting influence should remain voluntary. Furthermore, it is worthwhile to consider different viewer classes in live-streaming as well.

⁵⁸PC Gamer: *Twitch’s “Stream First” initiative integrates chat with a new wave of games*, <https://goo.gl/NH3a5F> (last accessed: 2018-07-07)

Chapter 7

General Conclusions

In this final chapter, we summarize the aspects that we investigated within this work. We then highlight the major theoretical, design and engineering contributions we made throughout the chapters. This is followed by a discussion of opportunities and challenges for future work: while some aspects have not been considered based on the thesis' scope, other aspects are now possible based on the contributions made.

7.1 Summary

We briefly considered the history and importance of games and play for humans through history. We saw that humans have an inherent desire to play that helps children to learn but also remains lifelong. Several contexts have appeared in which the fundamental concepts of games and play were used to make activities more engaging or game-like. This was done to harness the human ludic drive in these contexts. While we presented some examples, gamification and game live-streams were the focus of this thesis. We elaborated on the aspect that games have usually a voluntary nature and offer a freedom of choice for the players, elements that seem not to be focused on for these game-related contexts. This has also been shown to be problematic from a motivational theory point of view, as, for example, the autonomy systems offer is important for intrinsic motivation.

This led to the main question of this thesis: can we provide users with more influence at the runtime in gamification settings and game live-streams? Through our considerations of gamified self-sustaining systems (**RQ1**, Chapter 3), self-tailored gamification (**RQ2**, Chapter 4) and interactive game live-streams (**RQ3**, Chapter 5 and Chapter 6), we contributed to this question. In all these systems, users received more influence options:

Self-sustaining (crowd-based) systems improve their outcomes through users devoting effort to the system. The actual users thus have a fundamental influence on the systems' outcome and benefit directly from their efforts. In self-tailored gamification, we investigated customizable gamification approaches. Here, users can decide whether or not they want to use gamification at runtime. If they do want to use it, they can also combine game elements as they see fit and can further customize these elements to their needs. We called this "bottom-up" gamification. Thus, users have a fundamental influence on the motivational component of a system. Interactive game live-streams provide viewers (i.e., the consumers of the live-streams) with options to impact the stream in the presence of a streamer who is able to orchestrate them. We developed an understanding for how interactivity is perceived. We also considered several aspects for how to support it and provide enhanced interaction and communication channels. These offered viewers more influence in this kind of experience. We also considered streams in which the streamer is absent and viewers had full, but shared, control over the game being streamed. Here, we explored how the group of users can self-orchestrate and how we can support this further. Thus, in both experiences the users had fundamental influence on how the stream proceeds and could change the course of what they consume.

All these scenarios had a reciprocity effect: by exerting influence, the experiences these systems offered changed as well. In addition, in these contexts, we were able to investigate influence moderated only by the functions the systems offer (e.g., the available set of game elements in a "bottom-up" scenario), but also (loosely and tightly coupled) group settings, in which other humans also exert influence at the same time. Overall, we were able to show that more influence options are perceived positively in all these contexts, and added to the understanding of how to realize them through prototypes and user studies.

7.2 Major Contributions

We already illustrated the thesis' contributions to different Human-Computer Interaction sub-fields in Section 1.5. Furthermore, we highlighted the findings and contributions during the presentations and discussions of results throughout the chapters.

In this section, we will summarize the major ones by focusing on the theoretical, design and engineering contributions this thesis has made:

Theoretical contributions: We learned that the concept of self-sustaining systems already motivates people to put effort into a system. In addition, we saw that the usage of gamification can further increase the willingness to expend effort, in line with the related work (see Sections 1.2.1 and 2.4.1). We could also show that people derive implicit benefits while contributing in such systems: in our case, by having an implicit learning effect. These aspects add to the existing crowdsourcing literature.

We conducted several studies that showed the relevancy of more fundamental customization options in gamification for users at the system's runtime. We learned that users demand influence options on the game elements themselves but also on decisions as to how to combine them in a system. But we also saw that certain top-down elements in a user-driven gamification approach might also be interesting from a user's point of view. Furthermore, the studies showed that users are able to handle the offered choices and create gamification settings that have positive, qualitatively and quantitatively measurable effects. Our derived results can be used to inform such approaches. Overall, we introduced "bottom-up" gamification to the field and thus have added to the ongoing efforts toward tailoring gamified interventions (see Section 2.3).

We investigate how interactivity is perceived in game live-stream settings. In line with the related work (Section 2.5) we found that interactivity is indeed desired by live-stream consumers overall and that even viewers that do not want to actively use the interactive options appreciate them. We learned that there are different degrees of integrative and interactive aspects in typical live-streams. For many viewers, both start as soon as a streamer talks to them during the "performance". We also found that, if a streamer is present, his or her orchestration options need to be high, i.e., interactive options that alter the stream without giving a streamer veto options are not perceived well. This further underlines the exposed position of the streamer as performer. Our results gained towards interactive features can be used to inform novel concepts and features for live-streaming platforms to allow better experiences in the future. Overall, this adds to the body of knowledge on streamer-audience interaction (see Section 2.5.1).

Our investigation of shared game control settings in the live-streaming context revealed that these are, in contrast to single- or multiplayer games, not clearly appealing for viewers. Nonetheless, this thesis contributes an understanding on the group interactions in such settings, and provides methods to allow for groups' self-orchestration to support decision making in such scenarios. Besides considering different input aggregators, we also framed further means as game elements and provided insights into how these elements are used. Overall, to our knowledge, we are the first to have investigated this scientifically in the context of live-streaming (see Section 2.5.2).

Design contributions: With the two self-sustaining systems, we described examples for how to design such systems. The studies validated these concepts and provided further insights on what needs to be considered in the design process.

We provided two settings in which we used "bottom-up" gamification. We considered common game elements and how they can fit into the "bottom-up" idea. This is a design contribution, as these concepts can be easily adapted for other scenarios, including ones that are not in the task management or microtask solving context. Through the studies, we also provided a notion on how these elements were actually used, and provided further insights for how to realize "bottom-up" gamification.

We conceptualized and showed a system in which enhanced communication and interactive channels for live-streams are integrated. Concepts like *ballot box communication* [330] or direct interaction concepts in the streaming window were shown to be beneficial in our study. Overall, these concepts contribute to current live-streaming platform issues and the question of how to realize concepts that empower individuals here.

We presented input aggregators and features that allow self-orchestration in shared game control settings. We motivated the different approaches and how these are used and perceived by users. These concepts can also be used in other computer-mediated group scenarios.

Engineering contributions: Throughout the chapters, we presented several prototypes that realized the theoretical and design considerations. They serve as an engineering contribution made by this thesis, as they can be used as test-beds for further studies: *CrowdChess*, for example, was explicitly built to investigate more self-orchestration options in shared game control settings and options were already discussed in Section 6.3.2. All prototypes showed how the theoretical and design considerations can be realized within computational systems and thus are a valuable addition to the overall contribution of this thesis. All prototypes were evaluated: while the *Trash Game* was only evaluated in its conceptual state (see Section 3.3.3), all other prototypes were tested in scenarios that were either “in the wild” studies (e.g., *Helpstone*; see Section 5.3.2) or laboratory studies that mimic “in the wild” situations (e.g., *TPP++*; see Section 6.2.2). This shows that these prototypes were high-fidelity [252] and robust enough for such uses.

7.3 Future Work

With our focus on gamification and game live-streams as game-related sub-fields, some aspects were not considered in this thesis and are now left for future work. Additionally, based on our findings, new opportunities and challenges arise that can now be addressed. While we have reported future work options already throughout the chapters were reasonable, to conclude this thesis, we will summarize main aspects subsequently:

User influence in other game-related sub-fields: In Section 1.2, we presented game-related sub-fields that were not considered in this thesis. They can also be approached with a similar methodology as presented in this work in the future. Through this, it can be investigated how users can be empowered here as well. Based on our findings, we see this as reasonable to do in the context of serious games. Here, similar to the self-sustaining systems, these are made *for* the users: when they can adapt the game to their needs, it is likely that they can maximize what they get out of the system. On the other hand, games with a purpose are interesting to adapt as well, as a higher user engagement in these systems would help to more quickly achieve the actual goal these systems pursue.

User influence in games: It will also be interesting to investigate typical single- or multiplayer games with respect to the question of how individual user influence can be further increased. Today, based on developers that allow others to modify their games [236, 237], large modding communities exist that already shape games to their needs and make their modifications accessible to others [219]. Poretski and Arazy [236] showed that firms that encourage modding of a game (for example, through providing access to modding tools) attract many people who voluntarily spend hours creating modifications (“mods”) (as was also highlighted by Postigo [237]), also positively affecting sales. This happens because even non-programmers (i.e., normal players), can now change the game through mods as they like. Nonetheless, non-programmers typically cannot realize what they have in mind, as this would require programming skills. This is similar to a restricted set of “bottom-up” game elements, in which users also cannot create completely new mechanics. Thus, the question arises, whether games can also be made more customizable for this group as well, especially as this would provide valuable insights for the related aspects of this thesis (as the same patterns could be used for adding new mechanics in the contexts considered by the thesis). An interesting approach here was *please be nice*⁵⁹, in which a simple game was presented and the first to beat the game in an iteration could always suggest a new feature which was then implemented. It might be valuable to consider a similar setting in the future as well (potentially enhanced with the findings we made for group interactions). Based on what we learned, especially in the shared game control settings, it would also be interesting to create a game which is actually designed from scratch to account for this unique scenario. In our considerations, we utilized existing games (*Pokémon Red*, or *chess*) and retrofitted these. Potentially, the perception of such settings would improve if fundamental new game mechanics were implemented into them.

Combination of contexts: Based on our findings, it is reasonable to also investigate combinations of contexts. For example, gamified self-sustaining systems in this thesis offered only a “top-down” gamification. Using “bottom-up” gamification here, and studying its effects, seems reasonable. For example, *ExpenseControl* could be used with “bottom-up” concepts being introduced. As we showed how the “bottom-up” part can be realized with the *BU-Microtasks Platform*, this would be easy to achieve. Another option (see Section 6.4) is to combine the concepts of input aggregation and individual user influence through game elements into the typical live-streaming experiences when a streamer is present.

Sources of motivation: Considering the result that “bottom-up” gamification has positive effects, the question arises where these originate from. This is of similar importance for the other contexts. Are users more engaged because, through the influence options, they can set up systems optimally from a psychological point of view? Or is it simply the choice given in these systems that helps, even if users are not creating an optimal experience for themselves? We started investigating this in Section 4.5 and found that users do not select game elements

⁵⁹<http://pleasebenice.aran-koning.com> (last accessed: 2018-07-07)

in a “bottom-up” scenario as suggested by their player type or personality traits. Nonetheless, it is still an open question whether it is the choice or other factors that guide the selection. Based on our findings, it is easy to move forward in this direction: the study setting in Section 4.5 could be extended by presenting users with their gamification concepts implemented in a system. With this, it could be evaluated if these are then still motivational for them (without relying only on self-reported data). Additionally, to learn about the factor of choice, in a study users could be given the choice whether they want to use a (“top-down”) gamified approach. This could be compared to a group that receives no choice while being in the gamified condition, and others in a baseline condition without gamification. The latter would help to learn whether people who decided to (or not to) use gamification performed better than those that had no choice and were in the baseline condition. Based on Section 2.2.3, this can be assumed, but it requires validation.

Understanding individual differences in user autonomy: While it is already known that there are different player types in gamification (see Section 2.3), we found further indications that these are also present in the live-streaming context as features were perceived differently in our online study (see Section 5.2.2). We also learned that not all users were equally interested in exerting direct influence on the stream’s course, not only in this online study but also with *Helpstone* (see Section 5.3) and in *TPP++* (see Section 6.2). Understanding the individual differences for user autonomy in gamification and game live-streams is worthwhile to consider next. This will help to create systems that offer more customization options and would allow the creation of systems that recommend suitable elements to these users first, i.e., making the range of options more usable. This would also further reduce the effort that customization requires of users. In addition, this understanding can also inform personalization approaches as well.

Testing on a larger scale: Our prototypes were often evaluated only with a small number of participants. To increase the external validity, similar studies as done and reported in this thesis could be conducted with more participants. Our approaches always had a large user base in mind and implement mechanisms that are intended for this: either to harness the power of groups (e.g., the self-sustaining systems in Chapter 3) or to mitigate negative effects of a larger user base (e.g., *Helpstone* introduced mechanisms to handle information overload; see Section 5.3.1). Nonetheless, because of the nature of our studies, the number of participating users was not particularly large, leaving these validations open for future work. Based on the system concepts, the platforms can easily be used with more participants. Here, it will be interesting to learn how the perception of specific aspects changes when more users are active in parallel. For example, we hypothesized that the aggregators in the shared game control settings (see Chapter 6) might be perceived differently (e.g., as the aggregation benefits become more obvious) when the group sizes are larger.

List of Figures

1.1	<i>Chess</i> as an example of a game impacted by technology	2
1.2	A historical comparison of games	3
1.3	Profile view excerpt from the social network <i>LinkedIn</i>	7
1.4	Excerpt from the Q&A site <i>Stack Overflow</i>	7
1.5	Example of a game live-stream on <i>Twitch</i>	8
1.6	General schematic of reciprocity in a system	13
1.7	Overview of the thesis' goals	14
1.8	Structure of this thesis	16
2.1	User interface parts of <i>Cafe Flour Sack</i> in relation to rewards	38
2.2	The game creation wizard in the <i>Games for Crowds</i> platform	40
2.3	A chat example from <i>Chorus</i>	45
2.4	<i>VizWiz</i> example and <i>ESP game</i> interface	46
2.5	Examples of OCR-related tools	48
2.6	Examples of robot movements with different input aggregators . .	50
2.7	Alternative streaming tools	58
2.8	Examples of giving the audience more influence	60
2.9	Examples of shared game control	63
2.10	The game <i>Shairit</i>	65
2.11	Example of <i>Twitch Plays Pokémon</i>	67
3.1	Instantiated schematic of reciprocity in self-sustaining systems . .	72
3.2	Workflow of <i>ExpenseControl</i>	75
3.3	<i>ExpenseControl</i> views	76
3.4	Image processing in <i>ExpenseControl</i>	77
3.5	Microtask types in <i>ExpenseControl</i>	79
3.6	Game elements used in <i>ExpenseControl</i>	80
3.7	Error rates between baseline and <i>ExpenseControl</i> algorithms	84
3.8	Solved microtasks for each participant/week	86
3.9	Examples of augmented trash cans	88
3.10	Example pictures for every considered waste category	91
3.11	Example classification task of the questionnaire	92
3.12	Error rates of individuals and aggregation algorithms	96
3.13	An example public trash can for waste separation	100
3.14	Trash can prototype of the <i>Trash Game</i>	101
3.15	Classification screen of the app after a decision	103
4.1	Instantiated schematic of reciprocity in self-tailored gamification .	110
4.2	<i>BU-ToDo</i> main task user interface	122
4.3	Game element configuration screens in <i>BU-ToDo</i>	123
4.4	"No gamification" interface of the <i>BU-Microtasks Platform</i>	131
4.5	"Top-down" gamification interface of the <i>BU-Microtasks Platform</i> .	132
4.6	"Bottom-up" gamification interface of the <i>BU-Microtasks Platform</i> .	135

5.1	Instantiated schematic of reciprocity in interactive live-streams . .	162
5.2	The game <i>Superfight</i> on <i>Twitch</i>	164
5.3	Setup of the <i>B.E.A.R.D.S.</i> pen & paper session	165
5.4	<i>Helpstone</i> from a viewer's view	188
5.5	<i>Helpstone</i> from a streamer's view	188
5.6	Enlarged widgets of a viewer's view	189
5.7	Enlarged widgets of a viewer's and streamer's view	190
5.8	Number of interactions done with <i>Helpstone</i> per viewer	194
6.1	Instantiated schematic of reciprocity in shared game control . . .	200
6.2	Maximum parallel viewer count per day for <i>TP</i> channels	201
6.3	The <i>TPP++</i> interface	202
6.4	The different aggregators and their uptime in the <i>TPP++</i> study .	207
6.5	A screenshot of <i>CrowdChess</i> running on <i>Mixer</i>	212
6.6	The user interface of <i>CrowdChess</i>	216
6.7	The explanation shown for the EC and EL aggregators	217

List of Tables

3.1	Fun and Engagement answers in pre- and post-questionnaire . . .	85
3.2	Overview of solved tasks per user/week	85
3.3	Errors per waste category and a priori assigned difficulty level . .	94
3.4	Confusion matrix showing the aggregated classification results . .	94
4.1	The scenarios used in the a priori online study	115
4.2	The requirements derived from the a priori online study	117
4.3	Game element variation across all scenarios and participants . . .	119
4.4	Game element usage in the <i>BU-ToDo</i> study	127
4.5	Available game configurations in the selective conditions of the <i>BU-Microtasks Platform</i>	133
4.6	Number of solved tasks in the <i>BU-Microtasks Platform</i> conditions .	138
4.7	Results of the <i>IMI</i> in the <i>BU-Microtasks Platform</i> conditions	140
4.8	The scenarios used in the gamification concept study	144
4.9	The game elements found in the gamification concept design study	148
4.10	Significantly different element suggestions between scenarios in the gamification concept design study	151
4.11	Scenario-wise comparison of the GES mean answers of those who did (not) use the element in their concept	152
4.12	Correlations between GES and RE with <i>Hexad</i> user types	154
4.13	Correlations between GES and RE with the <i>Big Five</i>	155
4.14	Correlations between <i>Hexad</i> user types and game elements found in [306]	156
5.1	Example polls used in the <i>Rocket Beans TV</i> pen & paper format . .	167
5.2	Live-streamed elements that were found through free text fields that were not integrated in our element set	172
5.3	Elements related to the streaming experience or general features .	173
5.4	Elements that allow viewers to impact the stream or streamer . .	174
5.5	Elements related to the streamer's behavior	174
5.6	Elements related to the screen/audio composition of the stream .	175
5.7	Motivation statements used in the online questionnaire	180
5.8	Number of interactions with <i>Helpstone</i> per match and element . .	192
5.9	Relevance statements for the different <i>Helpstone</i> elements	193
5.10	Perception statements for the different <i>Helpstone</i> elements	193
6.1	The aggregators offered in <i>TPP++</i>	203
6.2	Players in our <i>CrowdChess</i> study	219
6.3	Aggregator performances based on all board positions and all move suggestions	222
6.4	Aggregator performances while active in a turn	223

Bibliography

- [1] Clark C. Abt. 1987. *Serious Games*. University Press of America.
- [2] Judith Ackermann and Marc Juchems. 2017. Twitch Plays Pokémon als Kollektive Let's Play-Performance. In *Phänomen Let's Play-Video*. Springer, 119–131.
- [3] Emily S. Adler and Roger Clark. 2014. *An Invitation to Social Research: How It's Done*. Cengage Learning.
- [4] Malik Almaliki, Nan Jiang, Raian Ali and Fabiano Dalpiaz. 2014. Gamified Culture-Aware Feedback Acquisition. In *Proceedings of the 7th International Conference on Utility and Cloud Computing (UCC '14)*. IEEE, 624–625.
- [5] Maximilian Altmeyer, Pascal Lessel and Antonio Krüger. 2016. Expense Control: A Gamified, Semi-Automated, Crowd-Based Approach for Receipt Capturing. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, 31–42.
- [6] Teresa M. Amabile and Judith Gitomer. 1984. Children's Artistic Creativity: Effects of Choice in Task Materials. *Personality and Social Psychology Bulletin* 10 (2). SAGE, 209–215.
- [7] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg and Jure Leskovec. 2013. Steering User Behavior with Badges. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. International World Wide Web Conferences Steering Committee, 95–106.
- [8] Sky LaRell Anderson. 2017. Watching People is Not a Game: Interactive Online Corporeality, Twitch.tv and Videogame Streams. *The International Journal of Computer Game Research* 17 (1). Game Studies, 1–16.
- [9] Dubravka Balen-Letunić. 2014. An Overview of Prehistoric Toys. *Etnološka Istraživanja* 1 (18/19). Etnografski Muzej, 11–17.
- [10] Aaron Bangor, Philip Kortum and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4 (3). Usability Professionals' Association, 114–123.
- [11] Richard Bartle. 1996. Hearts, Clubs, Diamonds, Spades: Players Who Suit MUDs. *Journal of MUD Research* 1 (1). 1–26.

- [12] Chris Bateman, Rebecca Lowenhaupt and Lennart E. Nacke. 2011. Player Typology in Theory and Practice. In *Proceedings of the 6th Digital Games Research Association (DiGRA '11)*. DiGRA/Utrecht School of the Arts, 1–24.
- [13] Alberto Battocchi, Fabio Pianesi, Daniel Tomasini, Massimo Zancanaro, Gianluca Esposito, Paola Venuti, Ayelet Ben-Sasson, Eynat Gal and Patrice L. Weiss. 2009. Collaborative Puzzle Game: A Tabletop Interactive Game for Fostering Collaboration in Children with Autism Spectrum Disorders (ASD). In *Proceedings of the 4th ACM International Conference on Interactive Tabletops and Surfaces (ITS '09)*. ACM, 197–204.
- [14] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick and Robert E. Kraut. 2004. Using Social Psychology to Motivate Contributions to Online Communities. In *Proceedings of the 7th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '04)*. ACM, 212–221.
- [15] Victoria Bellotti, Brinda Dalal, Nathaniel Good, Peter Flynn, Daniel G. Bobrow and Nicolas Ducheneaut. 2004. What a To-Do: Studies of Task Management Towards the Design of a Personal Task List Manager. In *Proceedings of the 22nd Annual ACM Conference on Human Factors in Computing Systems (CHI '04)*. ACM, 735–742.
- [16] Roland Bénabou and Jean Tirole. 2003. Intrinsic and Extrinsic Motivation. *The Review of Economic Studies* 70 (3). Oxford University Press, 489–520.
- [17] Karl Bergström and Staffan Björk. 2014. The Case for Computer-Augmented Games – Using Computers to Support and Not Dictate Gameplay. *Transactions of the 12th Digital Games Research Association* 1 (3). ETC Press, 1–32.
- [18] Leonard Berkowitz and Edward Donnerstein. 1982. External Validity is More Than Skin Deep: Some Answers to Criticisms of Laboratory Experiments. *American Psychologist* 37 (3). APA, 245–257.
- [19] Michael S. Bernstein, Joel Brandt, Robert C. Miller and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-Powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, 33–42.
- [20] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White and Tom Yeh. 2010. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, 333–342.
- [21] Max V. Birk, Cheralyn Atkins, Jason T. Bowey and Regan L. Mandryk. 2016. Fostering Intrinsic Motivation Through Avatar Identification in Digital Games. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2982–2995.

- [22] Max V. Birk, Maximilian A. Friehs and Regan L. Mandryk. 2017. Age-Based Preferences and Player Experience: A Crowdsourced Cross-Sectional Study. In *Proceedings of the 4th Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, 157–170.
- [23] Max V. Birk and Regan L. Mandryk. 2018. Combating Attrition in Digital Self-Improvement Programs Using Avatar Customization. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 660:1–660:15.
- [24] Martin Böckle, Isabel Micheel, Markus Bick and Jasminko Novak. 2018. A Design Framework for Adaptive Gamification Applications. In *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS '18)*. ScholarSpace, 1227–1236.
- [25] Martin Böckle, Jasminko Novak and Markus Bick. 2017. Towards Adaptive Gamification: A Synthesis of Current Developments. In *Proceedings of the 25th European Conference on Information Systems (ECIS '17)*. AIS Electronic Library, 158–173.
- [26] James Bonner and Clinton J. Woodward. 2012. On Domain-Specific Decision Support Systems for e-Sports Strategy Games. In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12)*. ACM, 42–51.
- [27] Simone Borges, Vinicius Durelli, Helena Reis, Ig I. Bittencourt, Riichiro Mizoguchi and Seiji Isotani. 2017. Selecting Effective Influence Principles for Tailoring Gamification-Based Strategies to Player Roles. In *Proceedings of the 28th Brazilian Symposium on Computers in Education (SBIE '17)*. Sociedade Brasileira de Computação, 857–866.
- [28] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the 31st Annual ACM Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 2117–2126.
- [29] John Brooke. 1996. SUS: A Quick and Dirty Usability Scale. In *Usability Evaluation in Industry*. Taylor & Francis, 189–194.
- [30] Michael Buhrmester, Tracy Kwang and Samuel D. Gosling. 2011. Amazon's Mechanical Turk a New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6 (1). SAGE, 3–5.
- [31] Marc Busch, Elke E. Mattheiss, Wolfgang Hochleitner, Christina Hochleitner, Michael Lankes, Peter Fröhlich, Rita Orji and Manfred Tscheligi. 2016. Using Player Type Models for Personalized Game Design – An Empirical Investigation. *Interaction Design and Architecture(s) – IxD&A Journal* 28. 145–163.

- [32] Carrie J. Cai, Shamsi T. Iqbal and Jaime Teevan. 2016. Chain Reactions: The Impact of Order on Microtask Chains. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 3143–3154.
- [33] Roger Callois. 1960. *Die Spiele und die Menschen*. Schwab.
- [34] Taj Campbell, Brian Ngo and James Fogarty. 2008. Game Design Principles in Everyday Fitness Applications. In *Proceedings of the 11th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '08)*. ACM, 249–252.
- [35] Vincenzo I. Cassone. 2017. Through the Ludic Glass: A Cultural Genealogy of Gamification. In *Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek '17)*. ACM, 54–62.
- [36] Teresa Cerratto-Pargman, Chiara Rossitto and Louise Barkhuus. 2014. Understanding Audience Participation in an Interactive Theater Performance. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction (NordiCHI '14)*. ACM, 608–617.
- [37] Pablo Cesar and David Geerts. 2011. Understanding Social TV: A Survey. In *Proceedings of the Networked and Electronic Media Summit (NEM Summit '11)*. Eurescom GmbH, 94–99.
- [38] Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid Crowd-Machine Learning Classifiers. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '15)*. ACM, 600–611.
- [39] Justin Cheng, Cristian Danescu-Niculescu-Mizil and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM '15)*. AAAI, 61–70.
- [40] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 4061–4064.
- [41] Gifford Cheung and Jeff Huang. 2011. Starcraft from the Stands: Understanding the Game Spectator. In *Proceedings of the 29th Annual ACM Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 763–772.
- [42] Chao-Min Chiu, Meng-Hsiang Hsu and Eric T. G. Wang. 2006. Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories. *Decision Support Systems* 42 (3). Elsevier, 1872–1888.

- [43] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt and Julie A. Kientz. 2014. Understanding Quantified-Selfers' Practices in Collecting and Exploring Personal Data. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 1143–1152.
- [44] Otto Chrons and Sami Sundell. 2011. Digitalkoot: Making Old Archives Accessible Using Crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation (AAAIWS'11-11)*. AAAI Press, 20–25.
- [45] David Codish and Gilad Ravid. 2015. Detecting Playfulness in Educational Gamification Through Behavior Patterns. *IBM Journal of Research and Development* 59 (6). IBM, 1–13.
- [46] Rob Comber and Anja Thieme. 2013. Designing Beyond Habit: Opening Space for Improved Recycling and Food Waste Behaviors Through Processes of Persuasion, Social Influence and Aversive Affect. *Personal Ubiquitous Computing* 17 (6). Springer, 1197–1210.
- [47] Rob Comber and Anja Thieme. 2017. BinCam: Evaluating Persuasion at Multiple Scales. In *Behavior Change Research and Theory*. Elsevier, 181–194.
- [48] Rob Comber, Anja Thieme, Ashur Rafiev, Nick Taylor, Nicole Krämer and Patrick Olivier. 2013. BinCam: Designing for Engagement with Facebook for Behavior Change. In *Proceedings of the 14th International Conference on Human-Computer Interaction (INTERACT '13)*. Springer, 99–115.
- [49] Thomas M. Connolly, Elizabeth A. Boyle, Ewan MacArthur, Thomas Hainey and James M. Boyle. 2012. A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games. *Computers & Education* 59 (2). Elsevier, 661–686.
- [50] Norman L. Corah and Joseph Boffa. 1970. Perceived Control, Self-Observation, and Response to Aversive Stimulation. *Journal of Personality and Social Psychology* 16 (1). APA, 1–4.
- [51] Paul T. Costa and Robert R. McCrae. 2008. The Revised NEO Personality Inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment* 2 (2). SAGE, 179–198.
- [52] Franco Curmi, Maria Angela Ferrario, Jon Whittle and Florian Mueller. 2015. Crowdsourcing Synchronous Spectator Support: (Go On, Go On, You're the Best)^{N-1}. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 757–766.
- [53] Paul G. Curran. 2016. Methods for the Detection of Carelessly Invalid Responses in Survey Data. *Journal of Experimental Social Psychology* 66. Elsevier, 4–19.

- [54] Nils Dahlbäck, Arne Jönsson and Lars Ahrenberg. 1993. Wizard of Oz Studies – Why and How. *Knowledge-Based Systems* 6 (4). Elsevier, 258–266.
- [55] Henry A. Davidson. 2012. *A Short History of Chess*. Three Rivers Press.
- [56] Gabriel V. de la Cruz, Bei Peng, Walter S. Lasecki and Matthew E. Taylor. 2015. Towards Integrating Real-Time Crowd Advice with Reinforcement Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion (IUI Companion '15)*. ACM, 17–20.
- [57] Edward L. Deci, Richard Koestner and Richard M. Ryan. 2001. Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. *Review of Educational Research* 71 (1). SAGE, 1–27.
- [58] Edward L. Deci and Richard M. Ryan. 2000. The “What” and “Why” of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11 (4). Taylor & Francis, 227–268.
- [59] Edward L. Deci and Richard M. Ryan. 2003. Intrinsic Motivation Inventory. 1–12.
- [60] Jie Deng, Felix Cuadrado, Gareth Tyson and Steve Uhlig. 2015. Behind the Game: Exploring the Twitch Streaming Platform. In *Proceedings of the 14th International Workshop on Network and Systems Support for Games (NetGames '15)*. IEEE, 8:1–8:6.
- [61] Sebastian Deterding. 2014. Eudaimonic Design, or: Six Invitations to Rethink Gamification. In *Rethinking Gamification*. Meson Press, 305–323.
- [62] Sebastian Deterding. 2015. The Lens of Intrinsic Skill Atoms: A Method for Gameful Design. *Human-Computer Interaction* 30 (3-4). Taylor & Francis, 294–335.
- [63] Sebastian Deterding. 2016. Contextual Autonomy Support in Video Game Play: A Grounded Theory. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 3931–3943.
- [64] Sebastian Deterding, Dan Dixon, Rilla Khaled and Lennart E. Nacke. 2011. From Game Design Elements to Gamefulness: Defining Gamification. In *Proceedings of the 15th International Academic Mindtrek Conference (Academic-Mindtrek '11)*. ACM, 9–15.
- [65] Alan Dix, Janet E. Finlay, Gregory D. Abowd and Russell Beale. 2003. *Human-Computer Interaction (3rd Edition)*. Prentice-Hall, Inc.
- [66] Anhai Doan, Raghu Ramakrishnan and Alon Y. Halevy. 2011. Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM* 54 (4). ACM, 86–96.

- [67] John Downs, Frank Vetere, Steve Howard and Steve Loughnan. 2013. Measuring Audience Experience in Social Videogaming. In *Proceedings of the 25th Australian Computer-Human Interaction Conference (OzCHI '13)*. ACM, 217–220.
- [68] Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries and Padmini Srinivasan. 2012. Quality Through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, 871–880.
- [69] Katharina Emmerich and Maic Masuch. 2017. The Impact of Game Patterns on Player Experience and Social Interaction in Co-Located Multiplayer Games. In *Proceedings of the 4th Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, 411–422.
- [70] Daniel Esser, Klemens Muthmann and Daniel Schuster. 2013. Information Extraction Efficiency of Business Documents Captured with Smartphones and Tablets. In *Proceedings of the 13th ACM Symposium on Document Engineering (DocEng '13)*. ACM, 111–114.
- [71] Paula Estrella and Pablo Paliza. 2014. OCR Correction of Documents Generated During Argentina's National Reorganization Process. In *Proceedings of the 1st International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*. ACM, 119–123.
- [72] Federal Ministry for the Environment, Nature Conservation and Nuclear Safety. 2016. Waste Management in Germany 2016: Facts, Data, Diagrams. 1–36.
- [73] Lauren S. Ferro, Steffen P. Walz and Stefan Greuter. 2013. Towards Personalised, Gamified Systems: An Investigation into Game Design, Personality and Player Typologies. In *Proceedings of the 9th Australasian Conference on Interactive Entertainment (IE '13)*. ACM, 7:1–7:6.
- [74] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek and Nigel Shadbolt. 2015. Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, 333–343.
- [75] Gerhard Fischer. 2011. Understanding, Fostering, and Supporting Cultures of Participation. *Interactions* 18 (3). ACM, 42–53.
- [76] Gerhard Fischer and Eric Scharff. 2000. Meta-Design: Design for Designers. In *Proceedings of the 3rd Conference on Designing Interactive Systems (DIS '00)*. ACM, 396–405.

- [77] Daniel Fitton and Chigbo Onyinyechukwu. 2017. Exploring Enforced Collaborative Agreement in Gaming with Young People. In *Proceedings of the 31st International BCS Human Computer Interaction Conference (BCS-HCI '17)*. BCS Learning & Development Ltd., 58:1–58:10.
- [78] Brian J. Fogg. 2002. Persuasive Technology: Using Computers to Change What We Think and Do. *Ubiquity* 2002 (12). ACM, 5:89–5:120.
- [79] Colin Ford, Dan Gardner, Leah E. Horgan, Calvin Liu, A. M. Tsaasan, Bonnie Nardi and Jordan Rickman. 2017. Chat Speed OP PogChamp: Practices of Coherence in Massive Twitch Chat. In *Proceedings of the 35th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, 858–871.
- [80] Guo Freeman and Donghee Y. Wohn. 2017. eSports as an Emerging Research Context at CHI: Diverse Perspectives on Definitions. In *Proceedings of the 35th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, 1601–1608.
- [81] Gesa Friederichs-Büttner, Benjamin Walther-Franks and Rainer Malaka. 2012. An Unfinished Drama: Designing Participation for the Theatrical Dance Performance Parcival XX-XI. In *Proceedings of the 7th Conference on Designing Interactive Systems (DIS '12)*. ACM, 770–778.
- [82] Mathilde B. Friedländer. 2017. Streamer Motives and User-Generated Content on Social Live-Streaming Services. *Journal of Information Science Theory and Practice* 5 (1). KISTI, 65–84.
- [83] Enrico Gandolfi. 2016. To Watch or to Play, It Is in the Game: The Game Culture on Twitch.tv among Performers, Plays and Audiences. *Journal of Gaming & Virtual Worlds* 8 (1). Intellect, 63–82.
- [84] Rosemary Garris, Robert Ahlers and James E. Driskell. 2002. Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming* 33 (4). SAGE, 441–467.
- [85] Lalya Gaye and Atau Tanaka. 2011. Beyond Participation: Empowerment, Control and Ownership in Youth-Led Collaborative Design. In *Proceedings of the 8th ACM Conference on Creativity and Cognition (C&C '11)*. ACM, 335–336.
- [86] David Geerts. 2006. Comparing Voice Chat and Text Chat in a Communication Tool for Interactive Television. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction (NordiCHI '06)*. ACM, 461–464.
- [87] David Geiger and Martin Schader. 2014. Personalized Task Recommendation in Crowdsourcing Information Systems – Current State of the Art. *Decision Support Systems* 65. Elsevier, 3–16.

- [88] Kathrin M. Gerling, Conor Linehan, Ben Kirman, Michael R. Kalyn, Adam B. Evans and Kieran C. Hicks. 2016. Creating Wheelchair-Controlled Video Games: Challenges and Opportunities when Involving Young People with Mobility Impairments and Game Design Experts. *Journal of Human Computer Studies* 94. Elsevier, 64–73.
- [89] Ken Goldberg, Billy Chen, Rory Solomon, Steve Bui, Bobak Farzin, Jacob Heitler, Derek Poon and Gordon Smith. 2000. Collaborative Teleoperation via the Internet. In *Proceedings of the 17th IEEE International Conference on Robotics and Automation (ICRA '00)*. IEEE, 2019–2024.
- [90] Ken Goldberg, Dezhen Song and Anthony Levandowski. 2003. Collaborative Teleoperation Using Networked Spatial Dynamic Voting. *Proceedings of the IEEE* 91 (3). IEEE, 430–439.
- [91] Lewis R. Goldberg. 1993. The Structure of Phenotypic Personality Traits. *American Psychologist* 48 (1). APA, 26–34.
- [92] Carina S. González, Pedro Toledo and Vanesa Muñoz. 2016. Enhancing the Engagement of Intelligent Tutorial Systems Through Personalization of Gamification. *International Journal of Engineering Education* 32 (1). Tempus Publications, 532–541.
- [93] Samuel D. Gosling, Simine Vazire, Sanjay Srivastava and Oliver P. John. 2004. Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions about Internet Questionnaires. *American Psychologist* 59 (2). APA, 93–104.
- [94] Samuel Greengard. 2010. Tracking Garbage. *Communications of the ACM* 53 (3). ACM, 19–20.
- [95] Daniel Gros, Brigitta Wanner, Anna Hackenholt, Piotr Zawadzki and Kathrin Knautz. 2017. World of Streaming. Motivation and Gratification on Twitch. In *Proceedings of the 9th International Conference on Social Computing and Social Media (SCSM '17)*. Springer, 44–57.
- [96] Anton Gustafsson, Cecilia Katzeff and Magnus Bang. 2010. Evaluation of a Pervasive Game for Domestic Energy Engagement Among Teenagers. *Computers in Entertainment* 7 (4). ACM, 54:1–54:19.
- [97] Carl Gutwin, Mutasem Barjawi and David Pinelle. 2016. The Emergence of High-Speed Interaction and Coordination in a (Formerly) Turn-Based Groupware Game. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. ACM, 277–286.
- [98] Ido Guy, Anat Hashavit and Yaniv Corem. 2015. Games for Crowds: A Crowdsourcing Game Platform for the Enterprise. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '15)*. ACM, 1860–1871.

- [99] Lassi Haaranen. 2017. Programming as a Performance: Live-Streaming and Its Implications for Computer Science Education. In *Proceedings of the 22nd ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '17)*. ACM, 353–358.
- [100] Oliver L. Haimson and John C. Tang. 2017. What Makes Live Events Engaging on Facebook Live, Periscope, and Snapchat. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 48–60.
- [101] Lasse Hakulinen and Tapio Auvinen. 2014. The Effect of Gamification on Students with Different Achievement Goal Orientations. In *Proceedings of the 2nd International Conference on Teaching and Learning in Computing and Engineering (LaTiCE '14)*. IEEE, 9–16.
- [102] Shivashankar Halan, Brent Rossen, Juan Cendan and Benjamin Lok. 2010. High Score! – Motivation Strategies for User Participation in Virtual Human Development. In *Intelligent Virtual Agents. Lecture Notes in Computer Science*, Vol. 6356. Springer, 482–488.
- [103] Juho Hamari. 2017. Do Badges Increase User Activity? A Field Experiment on the Effects of Gamification. *Computers in Human Behavior* 71. Elsevier, 469–478.
- [104] Juho Hamari and Lauri Keronen. 2017. Why Do People Play Games? A Meta-Analysis. *International Journal of Information Management* 37 (3). Elsevier, 125–141.
- [105] Juho Hamari and Jonna Koivisto. 2013. Social Motivations to Use Gamification: An Empirical Study of Gamifying Exercise. In *ECIS Completed Research*. AIS Electronic Library, 105:1–105:12.
- [106] Juho Hamari, Jonna Koivisto and Harri Sarsa. 2014. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS '14)*. IEEE, 3025–3034.
- [107] Juho Hamari and Janne Tuunanen. 2014. Player Types: A Meta-Synthesis. *Transactions of the Digital Games Research Association* 1 (2). ETC Press, 29–52.
- [108] William A. Hamilton, Oliver Garretson and Andruid Kerne. 2014. Streaming on Twitch: Fostering Participatory Communities of Play Within Live Mixed Media. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 1315–1324.
- [109] William A. Hamilton, John C. Tang, Gina Venolia, Kori M. Inkpen, Jakob Zillner and Derek Huang. 2016. Rivulet: Exploring Participation in Live Events Through Multi-Stream Experiences. In *Proceedings of the 14th ACM International Conference on Interactive Experiences for TV and Online Video (TVX '16)*. ACM, 31–42.

- [110] Carrie Heeter, Brian Magerko, Ben Medler and Yu-Hao Lee. 2011. Impacts of Forced Serious Game Play on Vulnerable Subgroups. *International Journal of Gaming and Computer-Mediated Simulations* 3 (3). IGI Global, 34–53.
- [111] Thomas T. Hewett, Ronald Baecker, Stuart Card, Tom Carey, Jean Gasen, Marilyn Mantei, Gary Perlman, Gary Strong and William Verplank. 1992. *ACM SIGCHI Curricula for Human-Computer Interaction*. ACM.
- [112] Zorah Hilvert-Bruce, James T. Neill, Max Sjöblom and Juho Hamari. 2018. Social Motivations of Live-Streaming Viewer Engagement on Twitch. *Computers in Human Behavior* 84. Elsevier, 58–67.
- [113] Alexis Hiniker, Bongshin Lee, Julie A. Kientz and Jenny S. Radesky. 2018. Let's Play! Digital and Analog Play Between Preschoolers and Parents. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 659:1–659:13.
- [114] Rose Holley. 2009. How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine* 15 (3/4). D-Lib Magazine, 1–13.
- [115] Rose Holley. 2009. *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*. National Library of Australia.
- [116] Daniel Hoornweg and Perinaz Bhada-Tata. 2012. What a Waste: A Global Review of Solid Waste Management. *Urban Development Series* 1 (15). The World Bank Group, 1–116.
- [117] Jacob Hornik, Joseph Cherian, Michelle Madansky and Chem Narayana. 1995. Determinants of Recycling Behavior: A Synthesis of Research Results. *The Journal of Socio-Economics* 24 (1). Elsevier, 105–127.
- [118] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15 (9). SAGE, 1277–1288.
- [119] Chin-Lung Hsu and Hsi-Peng Lu. 2004. Why Do People Play On-Line Games? An Extended TAM with Social Influences and Flow Experience. *Information and Management* 41 (7). Elsevier, 853–868.
- [120] Hsi-Mei Hsu and Chen-Tung Chen. 1996. Aggregation of Fuzzy Opinions Under Group Decision Making. *Fuzzy Sets and Systems* 79 (3). Elsevier, 279–285.
- [121] Mu Hu, Mingli Zhang and Yu Wang. 2017. Why Do Audiences Choose to Keep Watching on Live Video Streaming Platforms? An Explanation of Dual Identification Framework. *Computers in Human Behavior* 75. Elsevier, 594–606.

- [122] Shih-Wen Huang, Pei-Fen Tu, Wai-Tat Fu and Mohammad Amanzadeh. 2013. Leveraging the Crowd to Improve Feature-Sentiment Analysis of User Reviews. In *Proceedings of the 18th International Conference on Intelligent User Interfaces (IUI '13)*. ACM, 3–14.
- [123] Johan Huizinga. 1949. *Homo Ludens: A Study of the Play Element in Culture*. Routledge and Kegan Paul.
- [124] Kai Huotari and Juho Hamari. 2012. Defining Gamification: A Service Marketing Perspective. In *Proceedings of the 16th International Academic Mindtrek Conference (AcademicMindtrek '12)*. ACM, 17–22.
- [125] Glena H. Iten, Sharon T. Steinemann and Klaus Opwis. 2018. Choosing to Help Monsters: A Mixed-Method Examination of Meaningful Choices in Narrative-Rich Games and Interactive Narratives. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 341:1–341:13.
- [126] Sheena S. Iyengar and Mark R. Lepper. 2000. When Choice is Demotivating: Can One Desire Too Much of a Good Thing? *Journal of Personality and Social Psychology* 79 (6). APA, 995–1006.
- [127] Jakob Nielsen. 2006. The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities.
- [128] Bill Janssen, Eric Saund, Eric Bier, Patricia Wall and Mary Ann Sprague. 2012. Receipts2Go: The Big World of Small Documents. In *Proceedings of the 12th ACM Symposium on Document Engineering (DocEng '12)*. ACM, 121–124.
- [129] Shih-Ping Jeng and Ching-I Teng. 2008. Personality and Motivations for Playing Online Games. *Social Behavior and Personality* 36 (8). Scientific Journal Publishers, 1053–1060.
- [130] Yuan Jia, Yikun Liu, Xing Yu and Stephen Volda. 2017. Designing Leaderboards for Gamification: Perceived Differences Based on User Ranking, Application Domain, and Personality Traits. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 1949–1960.
- [131] Yuan Jia, Bin Xu, Yamini Karanam and Stephen Volda. 2016. Personality-Targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2001–2013.
- [132] Allie Johns, Adam Galpin, Joanne Meredith and Maxine Glancy. 2016. I Kind of Had an Avatar Switch: The Role of the Self in Engagement with an Interactive TV Drama. In *Proceedings of the 14th ACM International Conference on Interactive Experiences for TV and Online Video (TVX '16)*. ACM, 77–82.

- [133] Eric J. Johnson, Suzanne B. Shu, Benedict G.C. Dellaert, Craig Fox, Daniel G. Goldstein, Gerald Häubl, Richard P. Larrick, John W. Payne, Ellen Peters, David Schkade, Brian Wansink and Elke U. Weber. 2012. Beyond Nudges: Tools of a Choice Architecture. *Marketing Letters* 23 (2). Springer, 487–504.
- [134] Atreyi Kankanhalli, Mahdiah Taher, Huseyin Cavusoglu and Seung Hyun Kim. 2012. Gamification: A New Paradigm for Online User Engagement. In *Proceedings of the 33rd International Conference on Information Systems (ICIS '12)*. ScholarBank, 3573–3582.
- [135] Anne M. Kanstrup. 2012. A Small Matter of Design: An Analysis of End Users As Designers. In *Proceedings of the 12th Participatory Design Conference (PDC '12)*. ACM, 109–118.
- [136] Dennis L. Kappen, Jens Johannsmeier and Lennart E. Nacke. 2013. Deconstructing 'Gamified' Task-Management Applications. In *Proceedings of the 1st International Conference on Gameful Design, Research, and Applications (Gamification '13)*. ACM, 139–142.
- [137] Dennis L. Kappen, Pejman Mirza-Babaei and Lennart E. Nacke. 2017. Gamification Through the Application of Motivational Affordances for Physical Activity Technology. In *Proceedings of the 4th Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, 5–18.
- [138] Yamini Karanam, Leslie Filko, Lindsay Kaser, Hanan Alotaibi, Elham Makhsoom and Stephen Volda. 2014. Motivational Affordances and Personality Types in Personal Informatics. In *Proceedings of the 16th International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, 79–82.
- [139] Mehdi Kaytoue, Arlei Silva, Loïc Cerf, Wagner Meira, Jr. and Chedy Raïssi. 2012. Watch Me Playing, I Am a Professional: A First Study on Video Game Live Streaming. In *Proceedings of the 21st International Conference on World Wide Web Companion (WWW '12 Companion)*. ACM, 1181–1188.
- [140] Frederic Kerber, Pascal Lessel, Maximilian Altmeyer, Annika Kaltenhauser, Christian Neurohr and Antonio Krüger. 2014. Towards a Novel Digital Household Account Book. In *Proceedings of the 32nd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, 1921–1926.
- [141] Rilla Khaled and Asimina Vasalou. 2014. Bridging Serious Games and Participatory Design. *International Journal of Child-Computer Interaction* 2 (2). Elsevier, 93–100.
- [142] Reza Khoshkangini, Giuseppe Valetto and Annapaola Marconi. 2017. Generating Personalized Challenges to Enhance the Persuasive Power of Gamification. In *Persuasive '17 Workshop on Personalization in Persuasive Technology*. 70–83.

- [143] Keunyeong Kim, Michael G. Schmierbach, Saraswathi Bellur, Mun-Young Chung, Julia D. Fraustino, Frank Dardis and Lee Ahern. 2015. Is It a Sense of Autonomy, Control, or Attachment? Exploring the Effects of In-Game Customization on Game Enjoyment. *Computers in Human Behavior* 48. Elsevier, 695–705.
- [144] Spiro Kioussis. 2002. Interactivity: A Concept Explication. *New Media & Society* 4 (3). SAGE, 355–383.
- [145] Ana C. T. Klock, Isabela Gasparini, Marcelo S. Pimenta and José Palazzo M. de Oliveira. 2015. “Everybody is Playing the Game, but Nobody’s Rules are the Same”: Towards Adaptation of Gamification Based on Users’ Characteristics. *Bulletin of the IEEE Technical Committee on Learning Technology* 17 (4). IEEE, 22–25.
- [146] Hsiu-Chia Ko and Wen-Ning Wu. 2017. Exploring the Determinants of Viewers’ Loyalty Toward Beauty YouTubers: A Parasocial Interaction Perspective. In *Proceedings of the 1st International Conference on Education and Multimedia Technology (ICEMT ’17)*. ACM, 81–86.
- [147] Masatomo Kobayashi, Shoma Arita, Toshinari Itoko, Shin Saito and Hironobu Takagi. 2015. Motivating Multi-Generational Crowd Workers in Social-Purpose Work. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW ’15)*. ACM, 1813–1824.
- [148] Jonna Koivisto and Juho Hamari. 2014. Demographic Differences in Perceived Benefits from Gamification. *Computers in Human Behavior* 35. Elsevier, 179–188.
- [149] Oliver Korn, Markus Funk, Stephan Abele, Thomas Hörz and Albrecht Schmidt. 2014. Context-Aware Assistive Systems at the Workplace: Analyzing the Effects of Projection and Gamification. In *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA ’14)*. ACM, 38:1–38:8.
- [150] Robert E. Kraut, Judith Olson, Mahzarin Banaji, Amy Bruckman, Jeffrey Cohen and Mick Couper. 2004. Psychological Research Online: Report of Board of Scientific Affairs’ Advisory Group on the Conduct of Research on the Internet. *American Psychologist* 59 (2). APA, 105–117.
- [151] Sarah Kuhn and Michael J. Muller. 1993. Participatory Design. *Communications of the ACM* 36 (6). ACM, 24–29.
- [152] Ulrich Kull. 1982. Biologische Grundlagen Menschlichen Verhaltens. *Universitas* 37. Heidelberger Lese-Zeiten Verlag, 183–189.

- [153] Janaki Kumar. 2013. Gamification at Work: Designing Engaging Business Software. In *Proceedings of the 2nd International Conference of Design, User Experience, and Usability (DUXU '13)*. Springer, 528–537.
- [154] Harris Kyriakou. 2015. Twitch Plays Pokémon: An Exploratory Analysis of Crowd Collaboration. 1–9.
- [155] Marc-André K. Lafrenière, Jérémie Verner-Filion and Robert J. Vallerand. 2012. Development and Validation of the Gaming Motivation Scale (GAMS). *Personality and Individual Differences* 53 (7). Elsevier, 827–831.
- [156] Ellen J. Langer. 1975. The Illusion of Control. *Journal of Personality and Social Psychology* 32 (2). APA, 311–328.
- [157] Ellen J. Langer and Judith Rodin. 1976. The Effects of Choice and Enhanced Personal Responsibility for the Aged: A Field Experiment in an Institutional Setting. *Journal of Personality and Social Psychology* 34 (2). APA, 191–198.
- [158] Walter S. Lasecki, Christopher Homan and Jeffrey P. Bigham. 2014. Architecting Real-Time Crowd-Powered Systems. *Human Computation* 1 (1). Open Journal Systems, 67–93.
- [159] Walter S. Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P. Bigham and Michael S. Bernstein. 2015. Apparition: Crowdsourced User Interfaces That Come to Life As You Sketch Them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 1925–1934.
- [160] Walter S. Lasecki, Kyle I. Murray, Samuel White, Robert C. Miller and Jeffrey P. Bigham. 2011. Real-Time Crowd Control of Existing Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, 23–32.
- [161] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F. Allen and Jeffrey P. Bigham. 2013. Chorus: A Crowd-Powered Conversational Assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, 151–162.
- [162] Jisoo Lee, Erin Walker, Winslow Burleson, Matthew Kay, Matthew Buman and Eric B. Hekler. 2017. Self-Experimentation for Behavior Change: Design and Formative Evaluation of Two Approaches. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 6837–6849.
- [163] Herbert M. Lefcourt. 1973. The Function of the Illusions of Control and Freedom. *American Psychologist* 28 (5). APA, 417–425.

- [164] Amanda Lenhart, Joseph Kahne, Ellen Middaugh, Alexandra Rankin Macgill, Chris Evans and Jessica Vitak. 2008. Teens, Video Games, and Civics: Teens’ Gaming Experiences Are Diverse and Include Significant Social Interaction and Civic Engagement. *Pew Internet & American Life Project*. ERIC, 1–64.
- [165] Pascal Lessel and Maximilian Altmeyer. 2017. Tabletop Game Meets Live-Streaming: Empowering the Audience. In *CHI PLAY ’17 Workshop on Augmented Tabletop Games*. 28–32.
- [166] Pascal Lessel, Maximilian Altmeyer and Antonio Krüger. 2015. Analysis of Recycling Capabilities of Individuals and Crowds to Encourage and Educate People to Separate Their Garbage Playfully. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI ’15)*. ACM, 1095–1104.
- [167] Pascal Lessel, Maximilian Altmeyer and Antonio Krüger. 2018. Users As Game Designers: Analyzing Gamification Concepts in a “Bottom-Up” Setting. In *Proceedings of the 22nd International Academic Mindtrek Conference (AcademicMindtrek ’18)*. ACM, 1–10.
- [168] Pascal Lessel, Maximilian Altmeyer and Antonio Krüger. 2018. Viewers’ Perception of Elements Used in Game Live-Streams. In *Proceedings of the 22nd International Academic Mindtrek Conference (AcademicMindtrek ’18)*. ACM, 59–68.
- [169] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff and Antonio Krüger. 2016. “Don’t Whip Me With Your Games”: Investigating “Bottom-Up” Gamification. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI ’16)*. ACM, 2026–2037.
- [170] Pascal Lessel, Maximilian Altmeyer, Marc Müller, Christian Wolff and Antonio Krüger. 2017. Measuring the Effect of “Bottom-Up” Gamification in a Microtask Setting. In *Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek ’17)*. ACM, 63–72.
- [171] Pascal Lessel, Michael Mauderer, Christian Wolff and Antonio Krüger. 2017. Let’s Play My Way: Investigating Audience Influence in User-Generated Gaming Live-Streams. In *Proceedings of the 15th ACM International Conference on Interactive Experiences for TV and Online Video (TVX ’17)*. ACM, 51–63.
- [172] Pascal Lessel, Alexander Vielhauer and Antonio Krüger. 2017. CrowdChess: A System to Investigate Shared Game Control in Live-Streams. In *Proceedings of the 4th Annual Symposium on Computer-Human Interaction in Play (CHI PLAY ’17)*. ACM, 389–400.

- [173] Pascal Lessel, Alexander Vielhauer and Antonio Krüger. 2017. Expanding Video Game Live-Streams with Enhanced Communication Channels: A Case Study. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 1571–1576.
- [174] Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10 (8). Nauka, 707–710.
- [175] Ian Li, Anind Dey and Jodi Forlizzi. 2010. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the 28th Annual ACM Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 557–566.
- [176] Andreas Lieberoth. 2015. Shallow Gamification: Testing Psychological Effects of Framing an Activity as a Game. *Games and Culture* 10 (3). SAGE, 229–248.
- [177] Edwin A. Locke and Gary P. Latham. 2002. Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey. *American Psychologist* 57 (9). APA, 705–717.
- [178] Anna Loparev, Walter S. Lasecki, Kyle I. Murray and Jeffrey P. Bigham. 2014. Introducing Shared Character Control to Existing Video Games. In *Proceedings of the 9th International Conference on the Foundations of Digital Games (FDG '14)*. 1–8.
- [179] Jan Lorenz, Heiko Rauhut, Frank Schweitzer and Dirk Helbing. 2011. How Social Influence Can Undermine the Wisdom of Crowd Effect. *Proceedings of the National Academy of Sciences of the United States of America* 108 (22). National Academy of Sciences, 9020–9025.
- [180] Danielle Lottridge, Frank Bentley, Matt Wheeler, Jason Lee, Janet Cheung, Katherine Ong and Cristy Rowley. 2017. Third-Wave Livestreaming: Teens' Long Form Selfie. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, 20:1–20:12.
- [181] Zhicong Lu, Haijun Xia, Seongkook Heo and Daniel J. Wigdor. 2018. You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 466:1–466:13.
- [182] Xiaojuan Ma and Nan Cao. 2017. Video-Based Evanescent, Anonymous, Asynchronous Social Interaction: Motivation and Adaption to Medium. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, 770–782.

- [183] Thomas W. Malone, Robert Laubacher and Chrysanthos Dellarocas. 2009. Harnessing Crowds: Mapping the Genome of Collective Intelligence. *MIT Sloan Research Papers* 1 (1). SSRN, 1–20.
- [184] Albert E. Mannes. 2009. Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science* 55 (8). INFORMS, 1267–1279.
- [185] Sampada Marathe and S. Shyam Sundar. 2011. What Drives Customization? Control or Identity?. In *Proceedings of the 29th Annual ACM Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 781–790.
- [186] Andrzej Marczewski. 2015. User Types. In *Even Ninja Monkeys Like to Play: Gamification, Game Thinking and Motivational Design*. CreateSpace Independent Publishing Platform, 65–80.
- [187] Michael Margel. 2014. Twitch Plays Pokemon: An Analysis of Social Dynamics in Crowdsourced Games. 1–9.
- [188] Winter A. Mason, Frederica R. Conrey and Eliot R. Smith. 2007. Situating Social Influence Processes: Dynamic, Multidirectional Flows of Influence Within Social Networks. *Personality and Social Psychology Review* 11 (3). SAGE, 279–300.
- [189] Winter A. Mason and Duncan J. Watts. 2009. Financial Incentives and the “Performance of Crowds”. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Human Computation (HCOMP '09)*. ACM, 77–85.
- [190] Elaine Massung, David Coyle, Kirsten F. Cater, Marc Jay and Chris Preist. 2013. Using Crowdsourcing to Support Pro-Environmental Community Activism. In *Proceedings of the 31st Annual ACM Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 371–380.
- [191] Yu Matsuura and Sachiko Kodama. 2017. Cheer Me!: A Video Game System Using Live Streaming Text Messages. In *Proceedings of the 14th International Conference on Advances in Computer Entertainment Technology (ACE '17)*. Springer, 311–317.
- [192] Dan Maynes-Aminzade, Randy Pausch and Steve Seitz. 2002. Techniques for Interactive Audience Participation. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02)*. IEEE, 1–6.
- [193] Athanasios Mazarakis. 2013. Like Diamonds in the Sky: How Feedback can Boost the Amount of Available Data for Learning Analytics. *International Journal of Technology Enhanced Learning* 5 (2). Inderscience Publishers Ltd, 107–116.

- [194] Athanasios Mazarakis. 2017. Gamification: Eine Experimentelle Untersuchung der Spielelemente Abzeichen und Story. In *Mensch und Computer 2017 – Tagungsband*. Gesellschaft für Informatik e.V., 3–14.
- [195] Jim McCambridge, Marijn De Bruin and John Witton. 2012. The Effects of Demand Characteristics on Research Participant Behaviours in Non-Laboratory Settings: A Systematic Review. *PloS One* 7 (6). Public Library of Science, 1–6.
- [196] Nicole McMahon, Peta Wyeth and Daniel Johnson. 2012. Personality and Player Types in Fallout New Vegas. In *Proceedings of the 4th International Conference on Fun and Games (FnG '12)*. ACM, 113–116.
- [197] Adam W. Meade and S. Bartholomew Craig. 2012. Identifying Careless Responses in Survey Data. *Psychological Methods* 17 (3). APA, 1–19.
- [198] Elisa D. Mekler, Florian Brühlmann, Klaus Opwis and Alexandre N. Tuch. 2013. Disassembling Gamification: The Effects of Points and Meaning on User Motivation and Performance. In *Proceedings of the 31st Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)*. ACM, 1137–1142.
- [199] Elisa D. Mekler, Florian Brühlmann, Alexandre N. Tuch and Klaus Opwis. 2017. Towards Understanding the Effects of Individual Gamification Elements on Intrinsic Motivation and Performance. *Computers in Human Behavior* 71. Elsevier, 525–534.
- [200] Matthew K. Miller, John C. Tang, Gina Venolia, Gerard Wilkinson and Kori M. Inkpen. 2017. Conversational Chat Circles: Being All Here Without Having to Hear It All. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 2394–2404.
- [201] Walter Mischel, Yuichi Shoda and Rodolfo Mendoza-Denton. 2002. Situation-Behavior Profiles as a Locus of Consistency in Personality. *Current Directions in Psychological Science* 11 (2). SAGE, 50–54.
- [202] Seonah Mok, Jaehwan Jeon, Monson H. Hayes and Joonki Paik. 2013. Participating Interface. In *SIGGRAPH Asia 2013 Art Gallery (SA '13)*. ACM, 24:1–24:4.
- [203] Ethan R. Mollick and Nancy Rothbard. 2013. Mandatory Fun: Gamification and the Impact of Games at Work. *The Wharton School Research Paper Series* 1 (1). SSRN, 1–54.
- [204] Baptiste Monterrat, Michel Desmarais, Elise Lavoué and Sébastien George. 2015. A Player Model for Adaptive Gamification in Learning Environments. In *Artificial Intelligence in Education*. Lecture Notes in Computer Science, Vol. 9112. Springer, 297–306.

- [205] Baptiste Monterrat, Elise Lavoué and Sébastien George. 2014. Toward an Adaptive Gamification System for Learning Environments. In *Proceedings of the 6th International Conference on Computer-Supported Education (CSEDU '14)*. Springer, 115–129.
- [206] Benedikt Morschheuser, Juho Hamari and Jonna Koivisto. 2016. Gamification in Crowdsourcing: A Review. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS '16)*. IEEE, 4375–4384.
- [207] Sean A. Munson and Sunny Consolvo. 2012. Exploring Goal-Setting, Rewards, Self-Monitoring, and Sharing to Motivate Physical Activity. In *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '12)*. IEEE, 25–32.
- [208] Ilya Musabirov, Denis Bulygin, Paul Okopny and Ksenia Konstantinova. 2018. Between an Arena and a Sports Bar: Online Chats of eSports Spectators. *arXiv:1801.02862* 1 (1). 1–6.
- [209] Ilya Musabirov, Denis Bulygin, Paul Okopny and Ksenia Konstantinova. 2018. Event-Driven Spectators' Communication in Massive eSports Online Chats. In *Proceedings of the 36th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '18)*. ACM, LBW564:1–LBW564:6.
- [210] Isabel Briggs Myers, Mary H. McCaulley and Robert Most. 1985. *MBTI Manual, a Guide to the Development and Use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.
- [211] Lennart E. Nacke. 2017. Games User Research and Gamification in Human-Computer Interaction. *XRDS* 24 (1). ACM, 48–51.
- [212] Lennart E. Nacke, Chris Bateman and Regan L. Mandryk. 2014. BrainHex: A Neurobiological Gamer Typology Survey. *Entertainment Computing* 5 (1). Elsevier, 55–62.
- [213] Lennart E. Nacke, Anna Cox, Regan L. Mandryk and Paul Cairns. 2016. SIGCHI Games: The Scope of Games and Play Research at CHI. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, 1088–1091.
- [214] Gustavo Nascimento, Manoel Ribeiro, Loïc Cerf, Natália Cesário, Mehdi Kaytoue, Chedy Raïssi, Thiago Vasconcelos and Wagner Meira. 2014. Modeling and Analyzing the Video Game Live-Streaming Community. In *Proceedings of the 9th Latin American Web Congress (LA-WEB '14)*. IEEE, 1–9.
- [215] Anton J. Nederhof. 1985. Methods of Coping With Social Desirability Bias: A Review. *European Journal of Social Psychology* 15 (3). Wiley Online Library, 263–280.

- [216] M. P. Nevetha and A. Baskar. 2015. Applications of Text Detection and Its Challenges: A Review. In *Proceedings of the 3rd International Symposium on Women in Computing and Informatics (WCI '15)*. ACM, 712–721.
- [217] Edward Newell and Derek Ruths. 2016. How One Microtask Affects Another. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 3155–3166.
- [218] Scott Nicholson. 2012. Strategies for Meaningful Gamification: Concepts behind Transformative Play and Participatory Museums. In *Proceedings of the 3rd International Academic Conference on Meaningful Play (Meaningful Play '12)*. 1–16.
- [219] Scott Nicholson. 2012. A User-Centered Theoretical Framework for Meaningful Gamification. In *Proceedings of the 8th International Conference on Games + Learning + Society (GLS '12)*. 1–7.
- [220] Scott Nicholson. 2015. A RECIPE for Meaningful Gamification. In *Gamification in Education and Business*. Springer, 1–20.
- [221] Jędrzej Olejniczak. 2015. A Linguistic Study of Language Variety Used on Twitch.tv: Descriptive and Corpus-Based Approaches. In *Proceedings of the 4th International Conference on Redefining Community in Intercultural Context (RCIC '15)*. Air Force Academy Publishing House, 329–344.
- [222] Rita Orji, Regan L. Mandryk and Julita Vassileva. 2017. Improving the Efficacy of Games for Change Using Personalization Models. *ACM Transactions on Computer-Human Interaction* 24 (5). ACM, 32:1–32:22.
- [223] Rita Orji, Regan L. Mandryk, Julita Vassileva and Kathrin M. Gerling. 2013. Tailoring Persuasive Health Games to Gamer Type. In *Proceedings of the 31st Annual ACM Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 2467–2476.
- [224] Rita Orji, Lennart E. Nacke and Chrysanne Di Marco. 2017. Towards Personality-Driven Persuasive Health Games and Gamified Systems. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 1015–1027.
- [225] Rita Orji, Kiemute Oyibo and Gustavo F. Tondello. 2017. A Comparison of System-Controlled and User-Controlled Personalization Approaches. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, 413–418.
- [226] Rita Orji, Gustavo F. Tondello and Lennart E. Nacke. 2018. Personalizing Persuasive Strategies in Gameful Systems to Gamification User Types. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 435:1–435:14.

- [227] Rita Orji, Julita Vassileva and Regan L. Mandryk. 2014. Modeling the Efficacy of Persuasive Strategies for Different Gamer Types in Serious Games for Health. *User Modeling and User-Adapted Interaction* 24 (5). Springer, 453–498.
- [228] Kiemute Oyibo, Rita Orji and Julita Vassileva. 2017. The Influence of Culture in the Effect of Age and Gender on Social Influence in Persuasive Technology. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, 47–52.
- [229] Rui Pan, Lyn Bartram and Carman Neustaedter. 2016. TwitchViz: A Visualization Tool for Twitch Chatrooms. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, 1959–1965.
- [230] Eric Paulos and Tom Jenkins. 2005. Urban Probes: Encountering Our Emerging Urban Atmospheres. In *Proceedings of the 23rd Annual ACM Conference on Human Factors in Computing Systems (CHI '05)*. ACM, 341–350.
- [231] Nicole Peever, Daniel Johnson and John Gardner. 2012. Personality & Video Game Genre Preferences. In *Proceedings of the 8th Australasian Conference on Interactive Entertainment (IE '12)*. ACM, 20:1–20:3.
- [232] Anthony J. Pellicone and June Ahn. 2017. The Game of Performing Play: Understanding Streaming As Cultural Production. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 4863–4874.
- [233] Wei Peng, Jih-Hsuan Lin, Karin A. Pfeiffer and Brian Winn. 2012. Need Satisfaction Supportive Game Features as Motivational Determinants: An Experimental Study of a Self-Determination Theory Guided Exergame. *Media Psychology* 15 (2). Taylor & Francis, 175–196.
- [234] Johannes Pfau, Jan David Smeddinck, Georg Volkmar, Nina Wenig and Rainer Malaka. 2018. Do You Think This is a Game? Contrasting a Serious Game with a Gamified Application for Health. In *Proceedings of the 36th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '18)*. ACM, LBW069:1–LBW069:6.
- [235] Philip M. Podsakoff and Dennis W. Organ. 1986. Self-Reports in Organizational Research: Problems and Prospects. *Journal of Management* 12 (4). SAGE, 531–544.
- [236] Lev Poretski and Ofer Arazy. 2017. Placing Value on Community Co-Creations: A Study of a Video Game 'Modding' Community. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, 480–491.

- [237] Hector Postigo. 2007. Of Mods and Modders: Chasing Down the Value of Fan-Based Digital Game Modifications. *Games and Culture* 2 (4). SAGE, 300–313.
- [238] Andrew K. Przybylski, C. Scott Rigby and Richard M. Ryan. 2010. A Motivational Model of Video Game Engagement. *Review of General Psychology* 14 (2). Educational Publishing Foundation, 154–166.
- [239] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran and Karl Aberer. 2013. An Evaluation of Aggregation Techniques in Crowdsourcing. In *Web Information Systems Engineering (WISE 2013)*. Lecture Notes in Computer Science, Vol. 8181. Springer, 1–15.
- [240] Sheizaf Rafaeli, Gilad Ravid and Vladimir Soroka. 2004. De-Lurking in Virtual Communities: A Social Communication Network Approach to Measuring the Effects of Social and Cultural Capital. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS '04)*. IEEE, 1–10.
- [241] Dennis Ramirez, Jenny Saucerman and Jeremy Dietmeier. 2014. Twitch Plays Pokemon: A Case Study in Big G Games. In *Proceedings of the 9th Digital Games Research Association (DiGRA '14)*. 1–10.
- [242] Beatrice Rammstedt and Oliver P. John. 2007. Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41 (1). Elsevier, 203–212.
- [243] Nancy Ratner and Jerome Bruner. 1978. Games, Social Exchange and the Acquisition of Language. *Journal of Child Language* 5 (3). Cambridge University Press, 391–401.
- [244] Stuart Reeves, Steve Benford, Claire O'Malley and Mike Fraser. 2005. Designing the Spectator Experience. In *Proceedings of the 23rd Annual ACM Conference on Human Factors in Computing Systems (CHI '05)*. ACM, 741–750.
- [245] Stuart Reeves, Christian Greiffenhagen, Martin Flintham, Steve Benford, Matt Adams, Ju Row Farr and Nicholas Tandavanti. 2015. I'd Hide You: Performing Live Broadcasting in Public. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 2573–2582.
- [246] Peter Reichl, Christian Löw, Svenja Schröder, Thomas Schmidt, Bernhard Schatzl, Valon Lushaj, Oliver Hödl, Florian Güldenpfennig and Christopher Widauer. 2016. The Salome Experience: Opera Live Streaming and Beyond. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, 728–737.
- [247] Inbal Reif, Florian Alt, Juan David Hincapié Ramos, Katerina Poteriyaykina and Johannes Wagner. 2010. Cleanly: Trashducation Urban System. In *Proceedings of the 28th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, 3511–3516.

- [248] Ganit Richter, Daphne R. Raban and Sheizaf Rafaeli. 2015. Studying Gamification: The Effect of Rewards and Incentives on Motivation. In *Gamification in Education and Business*. Springer, 21–46.
- [249] C. Scott Rigby and Richard M. Ryan. 2007. The Player Experience of Need Satisfaction (PENS) Model. *Immersyve Inc.*. 1–22.
- [250] John M. Roberts, Malcolm J. Arth and Robert R. Bush. 1959. Games in Culture. *American Anthropologist* 61 (4). Wiley Online Library, 597–605.
- [251] Raquel Robinson, Zachary Rubin, Elena Márquez Segura and Katherine Isbister. 2017. All the Feels: Designing a Tool That Reveals Streamers’ Biometrics to Spectators. In *Proceedings of the 12th International Conference on the Foundations of Digital Games (FDG ’17)*. ACM, 36:1–36:6.
- [252] Yvonne Rogers, Helen Sharp and Jenny Preece. 2011. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons.
- [253] Donald Roy. 1959. “Banana Time”: Job Satisfaction and Informal Interaction. *Human Organization* 18 (4). Society for Applied Anthropology, 158–168.
- [254] Marco C. Rozendaal, Bram A. L. Braat and Stephan A. G. Wensveen. 2010. Exploring Sociality and Engagement in Play Through Game-Control Distribution. *AI & Society* 25 (2). Springer, 193–201.
- [255] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25 (1). Elsevier, 54–67.
- [256] Richard M. Ryan and Edward L. Deci. 2002. Overview of Self-Determination Theory: An Organismic Dialectical Perspective. In *Handbook of Self-Determination Research*. Boydell & Brewer Ltd., 3–33.
- [257] Richard M. Ryan and Edward L. Deci. 2006. Self-Regulation and the Problem of Human Autonomy: Does Psychology Need Choice, Self-Determination, and Will? *Journal of Personality* 74 (6). Wiley Online Library, 1557–1586.
- [258] Richard M. Ryan, C. Scott Rigby and Andrew K. Przybylski. 2006. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion* 30 (4). Springer, 344–360.
- [259] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI ’15)*. ACM, 1621–1630.

- [260] Elliot Salisbury, Sebastian Stein and Sarvapali Ramchurn. 2015. Real-Time Opinion Aggregation Methods for Crowd Robotics. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. International Foundation for Autonomous Agents and Multiagent Systems, 841–849.
- [261] Elizabeth B.-N. Sanders. 2006. Design Serving People. *Cumulus Working Papers* 15. University of Art and Design Helsinki, 28–33.
- [262] Elizabeth B.-N. Sanders and Pieter Jan Stappers. 2008. Co-Creation and the New Landscapes of Design. *CoDesign* 4 (1). Taylor & Francis, 5–18.
- [263] Walt Scacchi. 2010. Computer Game Mods, Modders, Modding, and the Mod Scene. *First Monday* 15 (5). First Monday Editorial Group.
- [264] Jesse Schell. 2008. *The Art of Game Design: A Book of Lenses*. Morgan Kaufmann Publishers Inc.
- [265] Sofia Schöbel, Matthias Söllner and Jan Marco Leimeister. 2016. The Agony of Choice – Analyzing User Preferences Regarding Gamification Elements in Learning Management Systems. In *Proceedings of the 37th International Conference on Information Systems (ICIS '16)*. Association for Information Systems, 1–21.
- [266] Douglas Schuler and Aki Namioka. 1993. *Participatory Design: Principles and Practices*. CRC Press.
- [267] Barry Schwartz. 2000. Self-Determination: The Tyranny of Freedom. *American Psychologist* 55 (1). APA, 79–88.
- [268] Rainforest Scully-Blaker, Jason Begy, Mia Consalvo and Sarah Ganzon. 2017. Playing Along and Playing For on Twitch: Livestreaming from Tandem Play to Performance. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS '17)*. IEEE, 2026–2035.
- [269] Katie Seaborn and Deborah I. Fels. 2015. Gamification in Theory and Action: A Survey. *International Journal of Human-Computer Studies* 74. Elsevier, 14–31.
- [270] Joseph Seering, Robert E. Kraut and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, 111–125.
- [271] Joseph Seering, Saiph Savage, Michael Eagle, Joshua Churchin, Rachel Moeller, Jeffrey P. Bigham and Jessica Hammer. 2017. Audience Participation Games: Blurring the Line Between Player and Spectator. In *Proceedings of the 12th Conference on Designing Interactive Systems (DIS '17)*. ACM, 429–440.

- [272] Ali Shamel, Tim Althoff, Amin Saberi and Jure Leskovec. 2017. How Gamification Affects Physical Activity: Large-Scale Analysis of Walking Challenges in a Mobile Application. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, 455–463.
- [273] David A. Shamma, Elizabeth F. Churchill, Nikhil Bobb and Matt Fukuda. 2009. Spinning Online: A Case Study of Internet Broadcasting by DJs. In *Proceedings of the 4th International Conference on Communities and Technologies (C&T '09)*. ACM, 175–184.
- [274] Kennon M. Sheldon, Andrew J. Elliot, Youngmee Kim and Tim Kasser. 2001. What is Satisfying About Satisfying Events? Testing 10 Candidate Psychological Needs. *Journal of Personality and Social Psychology* 80 (2). APA, 325–339.
- [275] Zhinian Shen and Yuri Tijerino. 2012. Ontology-Based Automatic Receipt Accounting System. In *Proceedings of the International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT '12)*. IEEE, 236–239.
- [276] Kristin Siu and Mark O. Riedl. 2016. Reward Systems in Human Computation Games. In *Proceedings of the 3rd Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16)*. ACM, 266–275.
- [277] Max Sjöblom and Juho Hamari. 2016. Why Do People Watch Others Play Video Games? An Empirical Study on the Motivations of Twitch Users. *Computers in Human Behavior* 75. Elsevier, 985–996.
- [278] Max Sjöblom, Maria Törhönen, Juho Hamari and Joseph Macey. 2017. Content Structure is King: An Empirical Study on Gratifications, Game Genres and Content Type on Twitch. *Computers in Human Behavior* 73. Elsevier, 161–171.
- [279] Thomas Smith, Marianna Obrist and Peter Wright. 2013. Live-Streaming Changes the (Video) Game. In *Proceedings of the 11th European Conference on Interactive TV and Video (EuroITV '13)*. ACM, 131–138.
- [280] Erica L. Snow, Laura K. Allen, G. Tanner Jackson and Danielle S. McNamara. 2015. Spendency: Students' Propensity to Use System Currency. *International Journal of Artificial Intelligence in Education* 25 (3). Springer, 407–427.
- [281] Stephen Snow and Dhaval Vyas. 2015. Fixing the Alignment: An Exploration of Budgeting Practices in the Home. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, 2271–2276.

- [282] Fabius Steinberger, Ronald Schroeter, Marcus Foth and Daniel Johnson. 2017. Designing Gamified Applications That Make Safe Driving More Engaging. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 2826–2839.
- [283] Steve Stemler. 2001. An Overview of Content Analysis. *Practical Assessment, Research & Evaluation* 7 (17). 137–146.
- [284] Ezra Stotland and Arthur L. Blumenthal. 1964. The Reduction of Anxiety as a Result of the Expectation of Making a Choice. *Canadian Journal of Psychology* 18 (2). University of Toronto Press, 139–145.
- [285] S. Shyam Sundar. 2008. Self as Source: Agency and Customization in Interactive Media. In *Mediated Interpersonal Communication*. Routledge, 58–74.
- [286] S. Shyam Sundar and Sampada S. Marathe. 2010. Personalization Versus Customization: The Importance of Agency, Privacy, and Power Usage. *Human Communication Research* 36 (3). Blackwell Publishing Ltd, 298–322.
- [287] S. Shyam Sundar, Jeeyun Oh, Saraswathi Bellur, Haiyan Jia and Hyang-Sook Kim. 2012. Interactivity As Self-Expression: A Field Experiment with Customization and Blogging. In *Proceedings of the 30th Annual ACM Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 395–404.
- [288] S. Shyam Sundar, Qian Xu and Saraswathi Bellur. 2010. Designing Interactivity in Media Interfaces: A Communications Perspective. In *Proceedings of the 28th Annual ACM Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2247–2256.
- [289] James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor.
- [290] William B. Swann Jr. and Thane S. Pittman. 1977. Initiating Play Activity of Children: The Moderating Influence of Verbal Cues on Intrinsic Motivation. *Child Development* 48. JSTOR, 1128–1132.
- [291] Philipp Sykownik, Katharina Emmerich and Maic Masuch. 2017. Exploring Patterns of Shared Control in Digital Multiplayer Games. In *Proceedings of the 14th International Conference on Advances in Computer Entertainment Technology (ACE '17)*. Springer, 847–867.
- [292] John C. Tang, Gina Venolia and Kori M. Inkpen. 2016. Meerkat and Periscope: I Stream, You Stream, Apps Stream for Live Streams. In *Proceedings of the 34th Annual CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 4770–4780.
- [293] Nicholas T. Taylor. 2016. Now You're Playing With Audience Power: The Work of Watching Games. *Critical Studies in Media Communication* 33 (4). Routledge, 293–307.

- [294] Nicholas T. Taylor. 2016. Play to the Camera: Video Ethnography, Spectatorship, and e-Sports. *Convergence* 22 (2). SAGE, 115–130.
- [295] Jaime Teevan, Shamsi T. Iqbal and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*. ACM, 2657–2668.
- [296] Jaime Teevan, Daniel J. Liebling and Walter S. Lasecki. 2014. Selfsourcing Personal Tasks. In *Proceedings of the 32nd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, 2527–2532.
- [297] Burak S. Tekin and Stuart Reeves. 2017. Ways of Spectating: Unravelling Spectator Participation in Kinect Play. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17)*. ACM, 1558–1570.
- [298] Richard H. Thaler, Cass R. Sunstein and John P. Balz. 2014. Choice Architecture. In *The Behavioral Foundations of Public Policy*. Achorn International, 428–439.
- [299] Anja Thieme, Rob Comber, Julia Miebach, Jack Weeden, Nicole Krämer, Shaun Lawson and Patrick Olivier. 2012. “We’ve Bin Watching You”: Designing for Reflection and Social Persuasion to Promote Sustainable Lifestyles. In *Proceedings of the 30th Annual ACM Conference on Human Factors in Computing Systems (CHI '12)*. ACM, 2337–2346.
- [300] Jennifer Thom, David Millen and Joan DiMicco. 2012. Removing Gamification from an Enterprise SNS. In *Proceedings of the 15th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '12)*. ACM, 1067–1070.
- [301] Crispin Thurlow, Laura Lengel and Alice Tomic. 2004. *Computer Mediated Communication*. SAGE.
- [302] Rose E. Timlett and Ian D. Williams. 2008. Public Participation and Recycling Performance in England: A Comparison of Tools for Behaviour Change. *Resources, Conservation and Recycling* 52 (4). Elsevier, 622–634.
- [303] Gustavo F. Tondello, Alberto Mora and Lennart E. Nacke. 2017. Elements of Gameful Design Emerging from User Preferences. In *Proceedings of the 4th Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, 129–142.
- [304] Gustavo F. Tondello and Lennart E. Nacke. 2018. Towards Customizing Gameful Systems by Gameful Design Elements. In *Persuasive '18 Workshop on Personalization in Persuasive Technology*. 1–9.

- [305] Gustavo F. Tondello, Rita Orji and Lennart E. Nacke. 2017. Recommender Systems for Personalized Gamification. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, 425–430.
- [306] Gustavo F. Tondello, Rina R. Wehbe, Lisa Diamond, Marc Busch, Andrzej Marczewski and Lennart E. Nacke. 2016. The Gamification User Types Hexad Scale. In *Proceedings of the 3rd Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16)*. ACM, 229–243.
- [307] Takumi Toyama, Daniel Sonntag, Andreas Dengel, Takahiro Matsuda, Masakazu Iwamura and Koichi Kise. 2014. A Mixed Reality Head-Mounted Text Translation System Using Eye Gaze Input. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. ACM, 329–334.
- [308] Khai N. Truong, Thariq Shhipar and Daniel J. Wigdor. 2014. Slide to X: Unlocking the Potential of Smartphone Unlocking. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 3635–3644.
- [309] Anders Tychsen. 2006. Role Playing Games: Comparative Analysis Across Two Media Platforms. In *Proceedings of the 3rd Australasian Conference on Interactive Entertainment (IE '06)*. Murdoch University, 75–82.
- [310] Marian F. Ursu, Maureen Thomas, Ian Kegel, Doug Williams, Mika Tuomola, Inger Lindstedt, Terence Wright, Andra Leurdijk, Vilmos Zsombori, Julia Sussner, Ulf Myrestam and Nina Hall. 2008. Interactive TV Narratives: Opportunities, Progress, and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 4 (4). ACM, 25:1–25:39.
- [311] José Van Dijck. 2009. Users Like You? Theorizing Agency in User-Generated Content. *Media, Culture & Society* 31 (1). SAGE, 41–58.
- [312] John Vines, Rachel Clarke, Peter Wright, John McCarthy and Patrick Olivier. 2013. Configuring Participation: On How We Involve People in Design. In *Proceedings of the 31st Annual ACM Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 429–438.
- [313] Joanne Vining and Angela Ebreo. 1990. What Makes a Recycler? A Comparison of Recyclers and Nonrecyclers. *Environment and Behavior* 22 (1). SAGE, 55–73.
- [314] Luis von Ahn. 2006. Games with a Purpose. *Computer* 39 (6). IEEE, 92–94.
- [315] Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the 22nd Annual ACM Conference on Human Factors in Computing Systems (CHI '04)*. ACM, 319–326.

- [316] Luis von Ahn, Mihir Kedia and Manuel Blum. 2006. Verbosity: A Game for Collecting Common-Sense Facts. In *Proceedings of the 24th Annual ACM Conference on Human Factors in Computing Systems (CHI '06)*. ACM, 75–78.
- [317] Luis von Ahn, Ruoran Liu and Manuel Blum. 2006. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the 24th Annual ACM Conference on Human Factors in Computing Systems (CHI '06)*. ACM, 55–64.
- [318] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321 (5895). American Association for the Advancement of Science, 1465–1468.
- [319] Peter Vorderer, Silvia Knobloch and Holger Schramm. 2001. Does Entertainment Suffer from Interactivity? The Impact of Watching an Interactive TV Movie on Viewers' Experience of Entertainment. *Media Psychology* 3 (4). Taylor & Francis, 343–363.
- [320] Michael G. Wagner. 2006. On the Scientific Relevance of eSports. In *Proceedings of the 7th International Conference on Internet Computing (ICOMP '06)*. CSREA Press, 437–442.
- [321] Molly McLure Wasko and Samer Faraj. 2005. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly* 29 (1). Society for Information Management and The Management Information Systems Research Center, 35–57.
- [322] Bernard Weiner. 1985. An Attributional Theory of Achievement Motivation and Emotion. *Psychological Review* 92 (4). APA, 548–573.
- [323] Justin D. Weisz, Sara Kiesler, Hui Zhang, Yuqing Ren, Robert E. Kraut and Joseph A. Konstan. 2007. Watching Together: Integrating Text Chat with Video. In *Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems (CHI '07)*. ACM, 877–886.
- [324] Kevin Werbach. 2014. (Re)Defining Gamification: A Process Approach. In *Proceedings of the 9th International Conference on Persuasive Technology (PERSUASIVE '14)*. Springer, 266–272.
- [325] Carol M. Werner and Eeva Makela. 1998. Motivations and Behaviors that Support Recycling. *Journal of Environmental Psychology* 18 (4). Elsevier, 373–386.
- [326] Mark Whooley, Bernd Ploderer and Kathleen Gray. 2014. On the Integration of Self-Tracking Data Amongst Quantified Self Members. In *Proceedings of the 28th International BCS Human Computer Interaction Conference (BCS-HCI '14)*. BCS, 151–160.

- [327] Eric N. Wiebe, Allison Lamb, Megan Hardy and David Sharek. 2014. Measuring Engagement in Video Game-Based Environments: Investigation of the User Engagement Scale. *Computers in Human Behavior* 32. Elsevier, 123–132.
- [328] Matthias Wilde, Katrin Bätz, Anastassiya Kovaleva and Detleff Urhahne. 2009. Testing a Short Scale of Intrinsic Motivation. *Zeitschrift für Didaktik der Naturwissenschaften* 15. Springer, 31–35.
- [329] Donghee Yvette Wohn, Guo Freeman and Caitlin McLaughlin. 2018. Explaining Viewers’ Emotional, Instrumental, and Financial Support Provision for Live Streamers. In *Proceedings of the 36th Annual ACM Conference on Human Factors in Computing Systems (CHI ’18)*. ACM, 474:1–474:13.
- [330] Mu Xia, Yun Huang, Wenjing Duan and Andrew B. Whinston. 2009. Ballot Box Communication in Online Communities. *Communications of the ACM* 52 (9). ACM, 138–142.
- [331] Xiao Xiao, Michael S. Bernstein, Lining Yao, David Lakatos, Lauren Gust, Kojo Acquah and Hiroshi Ishii. 2011. PingPong++: Community Customization in Games and Entertainment. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE ’11)*. ACM, 1–6.
- [332] Yan Xu, Erika Shehan Poole, Andrew D. Miller, Elsa Eiriksdottir, Richard Catrambone and Elizabeth D. Mynatt. 2012. Designing Pervasive Health Games for Sustainability, Adaptability and Sociability. In *Proceedings of the 7th International Conference on the Foundations of Digital Games (FDG ’12)*. ACM, 49–56.
- [333] Cong Zhang and Jiangchuan Liu. 2015. On Crowdsourced Interactive Live Streaming: A Twitch.tv-Based Measurement Study. In *Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV ’15)*. ACM, 55–60.
- [334] Cong Zhang, Jiangchuan Liu and Haiyang Wang. 2016. Towards Hybrid Cloud-Assisted Crowdsourced Live Streaming: Measurement and Analysis. In *Proceedings of the 26th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV ’16)*. ACM, 1:1–1:6.
- [335] Guangyu Zhu, Timothy J. Bethea and Vikas Krishna. 2007. Extracting Relevant Named Entities for Automated Expense Reimbursement. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’07)*. ACM, 1004–1012.
- [336] Melissa Zlatow and Aisling Kelliher. 2007. Increasing Recycling Behaviors Through User-Centered Design. In *Proceedings of the 3rd Conference on Designing for User eXperiences (DUX ’07)*. ACM, 27:1–27:1.

- [337] Miron Zuckerman, Joseph Porac, Drew Lathin and Edward L. Deci. 1978. On the Importance of Self-Determination for Intrinsically-Motivated Behavior. *Personality and Social Psychology Bulletin* 4 (3). SAGE, 443–446.
- [338] Oren Zuckerman and Ayelet Gal-Oz. 2014. Deconstructing Gamification: Evaluating the Effectiveness of Continuous Measurement, Virtual Rewards, and Social Comparison for Promoting Physical Activity. *Personal Ubiquitous Computing* 18 (7). Springer, 1705–1719.