Current Challenges in HCI-Research: Quantifying Open Experiences, Warranting Data Quality, and Developing Standardized Measures

Inaugural Dissertation

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy to the Department of Psychology,
of the University of Basel

by

Serge Petralito
from Rothrist (AG), Switzerland

Basel, 2019

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel
edoc.unibas.ch

CURRENT CHALLENGES IN HCI-RESEARCH

Approved by the Department of Psychology

At the request of

Prof. Dr. Klaus Opwis (First Reviewer)
Prof. Dr. Jana Nikitin (Second Reviewer)

Basel, Switzerland, _____

_____

Prof. Dr. Alexander Grob (Dean)

Contents

CURRENT CHALLENGES IN HCI-RESEARCH

Abstract

The three manuscripts that make up this dissertation represent three challenges of modern human-computer interaction (HCI) research and provide new insights, strategies, and recommendations for other researchers in this domain. The relatively new and fast-moving field of UX-research as yet provides insufficient theoretical groundwork in certain areas of interest. The first manuscript depicts the way in which a mixed-method approach, including qualitative and quantitative strategies, was able to reveal new dimensions of interest in the domain of challenges and avatar death in player experiences, a field previously characterized by a lack of the theoretical frameworks needed to address certain phenomena. Recent research in HCI is further complicated by the increasing trend of online data collection, a method which is concerned to provide insufficient data quality and therefore prone to failed replications or false effects. The second manuscript therefore aimed at providing a systematic analysis of a crowdsourced sample and practical recommendations, applying various measures to detect inattentive behavior. Lastly, a lack of common conceptual definitions, including the according measuring instruments, imposes another challenge on UX-researchers. The third manuscript revolves around the development and validation of a measure for trust on the web, a domain which previously lacked common concepts and measures.

This cumulative dissertation is based on the following three manuscripts:

1. Petralito, S., Brühlmann, F., Iten, G., Mekler, E. D., Opwis, K. (2017). A Good Reason to Die: How Avatar Death and High Challenges Enable Positive Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 5087-5097). ACM.
2. Brühlmann, F., Petralito, S., Aeschbach, L. F., Opwis, K. (submitted). Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples. [Manuscript submitted to *Plos One*]
3. Brühlmann, F., Petralito, S., Rieser, D. C., Aeschbach, L. F., Opwis, K. (submitted). TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web. [Manuscript submitted to *International Journal of Human-Computer Studies*]

Introduction

Research in human-computer interaction (HCI) has experienced a recent shift from a predominant focus on usability, mainly evaluating the overall effectiveness, efficiency, and satisfaction of interactions, to user experience (UX) (Hassenzahl, 2008; Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). In addition to the rather pragmatic and traditional methods used to attain usability goals, UX research fosters an approach based on the subjective and emotional experiences of users (Laugwitz, Schrepp, & Held, 2006). The sudden appearance of UX as a new, albeit very broad and vague, notion in HCI has led to a new branch of research that for many years lacked a generally accepted definition of UX (Mirnig, Meschtscherjakov, Wurhofer, Meneweger, & Tscheligi, 2015). One reason for this was its association with various ambiguous concepts and theoretical models deriving from emotional, affective, experiential, hedonic, and aesthetic variables (Law et al., 2009).

In 2010, the International Organization for Standardization (ISO) released a definitive UX definition, stating that UX is "a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service", further noting that "UX includes all the user's emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors and accomplishments that occur before, during and after use." (International Organization for Standardization, 2010). Already in 2008, Hassenzahl criticized this soon to be released UX definition, stating that too many concepts were included and that while this might lead to general agreement among researchers, there was still no clarification of the term itself (Hassenzahl, 2008).

The problems introduced briefly above associated with the term "user experience" and its relatively recent introduction in the field of HCI are at the core of three challenges in modern HCI-research, which are represented by the three manuscripts in the present dissertation. The first challenge is exemplified by player experience research, which is a prime example of a domain in which the sole functionality in the sense of usability is not sufficient to explore the somewhat contradictory desires of video game players, who generally strive to win yet at the same time want to be challenged and to learn through failing (Juul, 2009). The lack of theoretical frameworks for explaining the role of high challenges and avatar death in player experience research complicate data collection, as new dimensions of interest have to be explored first and turned into quantifiable experiences by applying a mixed-method approach.

The second challenge revolves around the increasing attention paid to data quality in online surveys and the various concerns that go along with it, as indicated by the number of studies conducted on crowdsourced websites which has drastically increased over the last few years (Chandler & Shapiro, 2016; Gosling & Mason, 2015). Finally, the third challenge discussed in this dissertation stems mainly from an overabundance of definitions and a lack of a generally agreed theoretical framework for users' trust on the web and, thus, the absence of a common, validated, and reliable measure for this dimension (Y. Kim & Peterson, 2017). Extensive research on trust in numerous academic fields has led to a multitude of definitions (Seckler, Heinz, Forde, Tuch, & Opwis, 2015; Van der Werff, Real, & Lynn, 2018) and measures being used in preexisting studies (e.g., Bart, Shankar, Sultan, & Urban, 2005; Bhattacherjee, 2002; Cho, 2006; Corbitt, Thanasankit, & Yi, 2003; Flavián, Guinalíu, & Gurrea, 2006; Jarvenpaa, Tractinsky, & Saarinen, 1999; Lee & Turban, 2001; McKnight, Choudhury, & Kacmar, 2002a; Pavlou & Gefen, 2004), which ultimately lack comparability and applicability in different contexts, thus further imposing a challenge on other researchers. These three challenges will now be introduced in more detail before addressing them in the relevant manuscripts of this dissertation.

Challenge 1: From Open Answers to a Quantifiable Experience

The first challenge for HCI-research is highlighted by the first manuscript of this dissertation. This study explores the relationship between high challenges, avatar death, and positive player experience. The majority of all preexisting literature in this domain is based on the theory of flow by Csikszentmihalyi (1990) and the corresponding concept of a challenge-skill balance, which is seen as the standard psychological explanation for the effects of failure and challenges on the player (Juul, 2009). While a certain level of challenges and avatar death is needed for video games to be interesting (Juul, 2009; Klarkowski et al., 2016; Sherry, 2004; Sweetser & Wyeth, 2005) and to provide informative experiences through learning (Flynn-Jones, 2015; Juul, 2013), the common perception is that excessive challenges are detrimental for video game enjoyment (Klimmt, Blake, Hefner, Vorderer, & Roth, 2009; Schmierbach, Chung, Wu, & Kim, 2014) and need to be adjusted to the player's skill level through adaptive difficulty mechanics (Cechanowicz, Gutwin, Bateman, Mandryk, & Stavness, 2014; Hunicke, 2005; Prendinger, Puntumapon, & Madruga, 2016; Tan, Tan, & Tay, 2011; Yun, Trevino, Holtkamp, & Deng, 2010) in order to provide a balanced experience. In addition to this theoretical paradigm, stan-

dardized scales for measuring players' enjoyment (Oliver & Bartsch, 2010), sense of challenge-skill balance (Schmierbach et al., 2014), challenge (IJsselsteijn et al., 2008), and other important concepts allow numerous dimensions within this field to be quantified.

This level of knowledge would assume that research questions concerning challenge and enjoyment in video games could thus best be examined using quantitative strategies and experimental research designs. However, the increasing popularity of video games like *Dark Souls*, which are notorious for their excessive difficulty as well as their frequent and highly punishing avatar death mechanism, raises some questions concerning the previously mentioned challenge theories. The popularity of such games suggests that high challenges and avatar death in video games play a more sophisticated role than merely being a factor that needs to be balanced out for players' enjoyment. In light of universally agreed design conventions such as the challenge-skill balance, the study discussed in the first manuscript aimed for a better understanding of why some players enjoy a game they constantly struggle with and fail at, and the role avatar death and high challenges play in this context.

The challenge to answer these questions lies in the lack of theoretical knowledge provided by the existing literature. To further explore important dimensions in this domain, a mixed-method approach that included the quantitative analysis of qualitative open answer data was necessary – an extensive and time-consuming procedure which requires coders to be well trained before they score protocols (Allison, Okun, & Dutridge, 2002; Reja, Manfreda, Hlebec, & Vehovar, 2003). The exploration of participants' subjective experiences and interpretations is a standard research procedure whenever empirical and theoretical knowledge is scarce. In HCI-research, the critical incident method and thematic analysis protocol (Braun & Clarke, 2006; Flanagan, 1954) are popular ways to explore research dimensions by applying a mixed-method approach with open-ended questions (e.g., Bopp, Mekler, & Opwis, 2016; Seckler et al., 2015; Tuch, Schaik, & Hornbæk, 2016), which also allows for a subsequent quantitative analysis to assess frequencies, co-occurrences, and other patterns. The inclusion of open-ended questions is generally recommended for exploring attitudes or values in a certain experience (Allison et al., 2002; Esses & Maio, 2002), as there are several advantages associated with them in comparison to closed-ended measures. Participants provide answers in their own words and are therefore not constrained by predefined terms provided by the response categories of closed-ended items (Esses & Maio, 2002; Holland & Christian, 2009; Reja et al.,

2003; Tourangeau, Rips, & Rasinski, 2000). Thus, open-ended questions such as the critical incident method allow for responses containing research dimensions that would otherwise not be covered by a closed-ended scale related to the same topic (Allison et al., 2002), thus exceeding the predefined scale contents and diversifying the set of answers (Reja et al., 2003). The data gathered by means of open-ended questions can subsequently be subjected to direct quantitative measures as well as content analyses (Esses & Maio, 2002). The content analysis from the thematic analysis protocol (Braun & Clarke, 2006) allows for important categories to be identified in both deductively and inductively by progressively analyzing the qualitative material.

The first manuscript presented in this dissertation illustrates the way in which a mixed-method approach, as described above, is able to reveal previously underexplored dimensions for a positive player experience in a setting of high challenges and frequent avatar death. The section devoted to the first manuscript provides a brief summary of the theoretical background as well as the methods, results, and discussion contained in the first manuscript.

Challenge 2: Data Quality From Crowdsourced Online Samples

The second challenge of modern HCI-research discussed in this dissertation revolves around warranting good data quality from crowdsourced online samples. With the appearance of numerous online crowdsourcing platforms like *Amazon's Mechanical Turk* (*MTurk*) and *TurkPrime* over the last decade, online data collection has gained more popularity than ever before (Kan & Drummey, 2018). The vast increase and rapid spread of online sample use is explained by the many advantages it has over traditional ways of data collection, including drastically lower costs (De Winter, Kyriakidis, Dodou, & Happee, 2015), lower hurdles for participation through relative anonymity (Kan & Drummey, 2018), diverse samples (Gosling & Mason, 2015; Paolacci & Chandler, 2014), and efficiency (Casler, Bickel, & Hackett, 2013).

However, in view of the increasing popularity of online data collection, concerns have been raised in regard to the data quality and the representativeness of such samples. Recently, various works have been dedicated to analyzing the reasons, effects, detection and prevalence of deficient data quality in online samples that stem from participants' inattention, carelessness and other deceptive behavior (Dogan, 2018; Hauser & Schwarz, 2016; Kan & Drummey, 2018; Maniaci & Rogge, 2014;

McKay, Garcia, Clapper, & Shultz, 2018; Meade & Craig, 2012; Peer, Brandimarte, Samat, & Acquisti, 2017). Online samples especially seem to be prone to inattentive or careless behavior, as participants in online studies are generally unsupervised and complete their surveys in an uncontrolled setting (Cheung, Burns, Sinclair, & Sliter, 2017). The detrimental effects of deficient data quality on the results of studies are manifold, for example failed replications of significant effects (Oppenheimer, Meyvis, & Davidenko, 2009), identifying non-existent effects (Huang, Liu, & Bowling, 2015), failed manipulations, and lower internal consistency of validated scales (Maniaci & Rogge, 2014). Some authors have even gone so far as to claim that poor data quality from inattentive or careless online samples is one of the main reasons for the recent replication crisis in psychological research (Maniaci & Rogge, 2014). However, the prevalence of careless or inattentive behavior in crowdsourced online samples remains largely unknown, as previous studies have mostly analyzed other forms of online samples (Maniaci & Rogge, 2014; Meade & Craig, 2012), did not apply various measures for carelessness (Dogan, 2018; Hauser & Schwarz, 2016; Peer et al., 2017), or examined other forms of deceptive behavior (Kan & Drummey, 2018).

The aim of the study explicated in the second manuscript was thus to systematically analyze the data quality of a crowdsourced online sample, based on various methods and recommendations, to detect careless behavior (Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012), thus addressing the limited variety of methods used in preexisting research about carelessness on crowdsourcing platforms. The study further analyzes the task dependency of careless behavior. Accordingly, the second manuscript presents a summary of the theoretical background as well as the results and the recommendations that were made to help other researchers address the challenge of guaranteeing good data quality in their online studies.

Challenge 3: Applying Common Standardized and Validated Measures

The third challenge, which is exemplified by the third manuscript of this dissertation, emerges from a multitude of definitions for trust, distrust and trust on the web (Seckler et al., 2015; Van der Werff et al., 2018), while lacking general agreement concerning the dimensional concepts and thus also a corresponding standardized measure. As many subdomains in HCI-research are relatively new and fast-moving, different theoretical frameworks are applied, impeding the general applicability of the measures used in these studies. Despite trust being a widely discussed research dimension across a wide range of disciplines (Van der Werff et al., 2018), there is

still a lack of a common, validated, reliable, versatile, and easy-to-translate measure for trust on the web (Y. Kim & Peterson, 2017).

Although trust on the web, like trust in an offline context, usually comprises the three subdimensions of benevolence, integrity, and competence (McKnight, Choudhury, & Kacmar, 2002b), there are also numerous variations and much disagreement on these concepts. Some authors introduce further subdimensions such as predictability or value congruence (Dietz & Den Hartog, 2006; McKnight, Cummings, & Chervany, 1998), while others argue that they are already covered by the previously mentioned subdimensions (Van der Werff et al., 2018). Furthermore, as many technological aspects are involved in the web context, trust in this case additionally refers to a non-conscious trustee. Therefore, some authors argue for including subdimensions like performance, helpfulness, predictability, or functionality (McKnight, Carter, Thatcher, & Clay, 2011; Söllner, Pavlou, & Leimeister, 2013), which view trust as a factor that is determined more by functional or technical dimensions and less by value congruence and interpersonal expectations. However, some studies have also found that online customers sometimes treat advanced technological agents as conscious beings and therefore the traditional dimensions of trust might nonetheless fully apply to this context (Lankton & McKnight, 2011; Wang & Emurian, 2005). This discourse illustrates that a general understanding of a concept might also change over time owing to advancements in technology and thus further complicating the applicability of a common measurement.

The lack of a common measure for trust on the web can further be traced back to the variety of specific contexts in which trust on the web had to be measured. The majority of existing works measure trust by applying tailor-made questionnaires and scales, which require participants to respond using Likert-type scales about a specific website or situation (e.g., Bart et al., 2005; Bhattacherjee, 2002; Cho, 2006; Corbitt et al., 2003; Flavián et al., 2006; Jarvenpaa et al., 1999; Lee & Turban, 2001; McKnight et al., 2002a; Pavlou & Gefen, 2004). Some of these scales do not cover all subdimensions of trust (Cho, 2006) and applying these specific Likert-type items in a different context or language would require extensive rephrasing and translating, which could have a negative impact on the reliability and validity of a scale. Taken together, the lack of a common validated measure imposes a challenge on researchers, as existing scales are not necessarily suitable for new studies and may require the majority of scale items to be extensively rephrased.

The studies discussed in the third manuscript are thus aimed at addressing this challenge by developing a semantic differential for measuring trust on the web, as this type of measure holds several advantages over Likert-type scales (Verhagen, Van Den Hooff, & Meents, 2015). The theoretical background to these studies, their methods and results, as well as the potential of the semantic differential scale *Trust-Diff*, will be extensively discussed in the respective section of the third manuscript.

Each of the three aforementioned challenges of HCI-research will now be addressed in the subsequent main chapters of the three manuscripts for this dissertation: "Positive player experiences in a setting of high challenges", "Data quality from crowdsourcing platforms", and "Measuring trust on the web". Each chapter contains an introduction to the respective theoretical background as well as a summary of the methods, results and discussion in the respective manuscript. Please note that further details concerning the methodological procedures and results are available in the corresponding manuscripts in the appendix.

Positive Player Experiences in a Setting of High Challenges

The first manuscript of this dissertation depicts the exploration of new dimensions relevant to the domain of challenges in player experiences and how a mixed-method approach helped to find answers in regard to a quantifiable experience. Before summarizing the methods, results and discussion in this manuscript, the theoretical background of this domain will be introduced in order to fully grasp the initial position of this work.

In order to attain a positive player experience, challenge is unanimously seen as one of the most important factors contributing to this interactive experience. Juul (2009) describes the desires of video game players as somewhat contradictory, as although they strive to win, a game without challenge or failure would likely be perceived as shallow and uninteresting (Juul, 2009; Klarkowski et al., 2016; Sherry, 2004; Sweetser & Wyeth, 2005). Juul (2009) analyzed open statements relating to why unchallenging games are not perceived as being enjoyable and found that failure is far more than merely a contrast to winning: failure pushes the player to reconsider strategy and thereby adds content to the game. However, while challenge and competition have been found to generally increase enjoyment in some studies (Lazzaro, 2004; Ryan, Rigby, & Przybylski, 2006; Vorderer, Klimmt, & Ritterfeld, 2004), other works have found higher enjoyment ratings for games with easier difficulty settings and have concluded that excessive challenges may undermine enjoyment (Klimmt et al., 2009; Schmierbach et al., 2014). On a similar notion, various studies emphasize that excessive challenges may be frustrating and therefore argue in favor of a balanced experience (Cechanowicz et al., 2014; Hunicke, 2005; Prendinger et al., 2016; Tan et al., 2011; Yun et al., 2010).

Challenge-Skill Balance and the Theory of Flow

The contradictory desires of players and the need for a balanced experience are often addressed using the theory of flow (Csikszentmihalyi, 1990), which is seen as the standard psychological explanation for game failure and challenge (Juul, 2009). According to this theory, a balance between the challenges imposed by the game and the players' skill level needs to be maintained in order to reach an optimal experience, resulting in a state of flow (see Figure 1). Hence, if the demands imposed by the game are too high in regard to the players' skill level, the experience will lead to anxiety, whereas a game with challenges set too low will be perceived as

boring. Falstein (2005) refined the standard concept of the challenge-skill balance of Csikszentmihalyi (1990), since the original illustration suggests a steady and smooth increase in difficulty over time. Falstein (2005) however argued, as seen in Figure 1, that the game difficulty should increase in waves. This variety of a game being sometimes rather easy and sometimes rather hard leads to enjoyment because an irregular increase in difficulty makes it more likely that the player will experience both failure and success (Falstein, 2005; Juul, 2009).

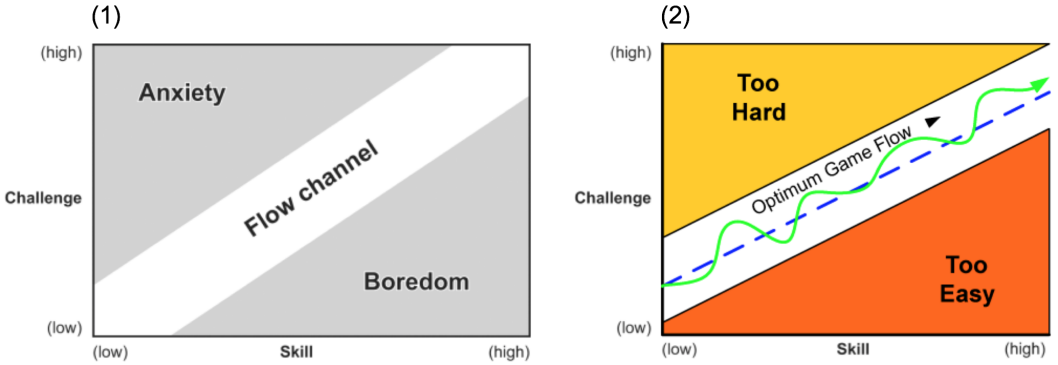


Figure 1. (1) Theory of flow by Csikszentmihalyi (1990). (2) Refined challenge-skill balance concept by Falstein (2005).

The majority of all current research surrounding challenge and failure in video games emphasizes the significance of an ideal challenge-skill balance based on the theory of flow (Aponte, Levieux, & Natkin, 2011; Bateman, Mandryk, Stach, & Gutwin, 2011; Gutwin, Rooke, Cockburn, Mandryk, & Lafreniere, 2016; Klarkowski et al., 2016; Ritterfeld, Cody, & Vorderer, 2009). Furthermore, an entire sub-branch of studies is devoted to the question of how to configure the difficulty of a game to a player's skill level individually through adaptive and dynamic difficulty adjustments in order to reach an ideal challenge-skill balance (Cechanowicz et al., 2014; Hunicke, 2005; Prendinger et al., 2016; Tan et al., 2011; Yun et al., 2010). These dynamic difficulty adjustments comprise various mechanics which ensure that less skilled players stay competitive in a game and may eventually succeed in achieving their goals. Common examples are the rubber band adjustment in racing games and auto-aiming features in combat games and first-person shooters (Bateman et al., 2011; Cechanowicz et al., 2014).

Difficulty and Enjoyment

In the literature, excessive challenges are generally portrayed as a negative factor for enjoyment (Klimmt et al., 2009; Schmierbach et al., 2014), which has to be throttled and adjusted to the player's skill level through adaptive difficulty mechanics (Cechanowicz et al., 2014; Hunicke, 2005; Prendinger et al., 2016; Tan et al., 2011; Yun et al., 2010). These mechanics have been proven to be preferred by novices in order to feel more competent, autonomous and related to other players (Cechanowicz et al., 2014). In addition, the inclusion of adaptive difficulty mechanics has been found to make games more enjoyable (Vicencio-Moreira, Mandryk, & Gutwin, 2015) and to support flow (J. Chen, 2007). Furthermore, ideally balanced games are associated with heightened positive affect, greater enjoyment as well as a heightened sense of autonomy and relatedness (Klarkowski et al., 2016).

Some studies not only emphasize the importance of a balanced experience; they argue even further that generally lowered difficulty settings are actually preferred by players. A study conducted by Schmierbach et al. (2014) examined the relationship between difficulty and enjoyment and the possible mediating roles of competence and challenge-skill balance. Their results demonstrated that playing the harder version of a game diminished the players' sense of competence by lowering their sense of challenge-skill balance, which in turn resulted in lower enjoyment ratings. Furthermore, avoiding avatar death has been found to be associated with higher levels of competence and flow (Jin, 2012). A study conducted by Klimmt et al. (2009) indicated that even experienced players reported greater enjoyment and satisfaction when playing a shooter game in easy mode compared to medium or hard settings.

Peng, Lin, Pfeiffer, and Winn (2012) demonstrated that adaptive difficulty primarily helps to increase enjoyment by reducing the challenges of the game. These findings can be put somewhat into perspective by other studies, indicating that only casual players enjoy lower difficulty settings, regardless of their skill level, while experienced players prefer challenges according to their abilities (Alexander, Sear, & Oikonomou, 2013) and that players prefer lower levels of difficulty mainly at the beginning of the game (Klimmt et al., 2009). To even out the challenge-skill balance in principle means to balance the occurrence of failure in a player's experience. Thus, the mechanics and role of failure or avatar death itself in the context of video games will be examined next.

Learning by Failing: The Role of Avatar Death

High challenges in regard to the player's skill level are likely to cause the player to fail and be removed from play, a game mechanism which is often referred to as avatar death and occurs in most genres of video games (Carter, Gibbs, & Wadley, 2013; Copcic, McKenzie, & Hobbs, 2013). Despite being a recurring element in games even well before computers existed (e.g. table-top games), the lack of research on the role of death in this context is surprising, considering its importance to the player and the immense success of the interactive entertainment industry (Copcic et al., 2013). Generally, avatar death as a mechanism imposes a penalty on the player, consisting of repetition and incremental progress towards mastery of a certain section of the game (Flynn-Jones, 2015). Similar to Juul's notion that failure pushes the player into rethinking their strategies (Juul, 2009), avatar death is often depicted as part of a learning process. Flynn-Jones (2015) describes in-game death and its accompanying loss of control as a recurring process of agency in terms of getting back control after death, repetition and ultimately mastery of a certain section of the game.

The motivation evolving from this mechanism stems from an informative learning process of how to overcome failure, which is critical to enjoyment (Flynn-Jones, 2015; Juul, 2013). Gee (2005) similarly stressed the notion of learning by failing but further stated that learning in a video game works best when challenges are pleasantly frustrating and the learner perceives himself as being at the outer edge but still within his range of competence. Thus, this theory suggests that if avatar death merely occurs to a certain extent, failure may be just an enjoyable experience you learn from, again pleading for an overall balanced approach to challenge.

The role of avatar death and its effects on the player can be further illuminated by examining how punishing the consequences resulting from failure are for the player. In MMOGs (Massively Multiplayer Online Games) there has been a gradual lessening of punishment by avatar death to the point where dying in the game is a non-event – an activity similar to other repeating occurrences as part of everyday life in the game's world (Klastrup, 2006). Although Gee (2005) argued that low consequences resulting from failure are a good thing to enable a fast learning process, Bartle (2004) noted that such a form of non-consequentiality may diminish the value of a player's acts, because this way every player can easily learn from past mistakes and if everyone is so easily a hero, then no one is (Copcic et al., 2013).

If avatar death in a game is meaningless then the game will soon be fully explored and become boring (Klastrup, 2006). On the other side of the gravity spectrum of punishments, two studies conducted by Carter et al. (2013) and Allison, Carter, and Gibbs (2015) examined the highly consequential death mechanism in the game *DayZ*, where every avatar death forces the player to restart the game from the very beginning – called "permadeath". The studies concluded that this consequentiality leads to a raised level of perceived investment (Carter et al., 2013) and that the awareness of risks imbues actions with meaning (Allison et al., 2015). While this meaningfulness stemmed from a pattern of smaller negative experiences, any achievement in the game in exchange was received as extremely positive as a contrast to the high stakes the game imposes on the player (Allison et al., 2015). Allison et al. (2015) further stated that negative and positive affect are not mutually exclusive: the positive experience is directly created by negative feelings of fear, anxiety and unpredictability. Avatar death may thus result in an immediate negative experience but ultimately leads to positive experiences if the players are able to achieve their goal.

Research Gap: The Enjoyment of Excessive Challenges

In conclusion, player experience research so far has associated excessive challenges with lower levels of enjoyment (Gutwin et al., 2016; Klimmt et al., 2009; Peng et al., 2012; Schmierbach et al., 2014), competence (Jin, 2012; Schmierbach et al., 2014), challenge-skill balance (Schmierbach et al., 2014) and flow (Jin, 2012) and thus argues that lower difficulty settings are to be preferred. Ideally balanced experiences are said to be key to positive player experiences as they provide the player with more enjoyment (Klarkowski et al., 2016; Vicencio-Moreira et al., 2015), competence (Cechanowicz et al., 2014), autonomy (Cechanowicz et al., 2014; Klarkowski et al., 2016), relatedness (Cechanowicz et al., 2014; Klarkowski et al., 2016), and flow (J. Chen, 2007). Although avatar death is a learning mechanism integral to enjoyment (Flynn-Jones, 2015; Juul, 2013), it should not undermine the players' feelings of competence and the consequences resulting from failure should be kept low (Gee, 2005). Highly punishing mechanisms are often viewed as a needlessly harsh penalty, especially if they occur in the wake of a small mistake and/or in the pursuit of a small reward, and they can diminish a player's overall enjoyment, motivation and progression (Copcic et al., 2013). These findings and statements however raise the question of why games such as the highly successful *Dark Souls* series are enjoyable

for many players. The *Dark Souls* games are notorious for their excessive difficulty as well as frequent and highly consequential avatar death, which stands in contrast to the majority of modern games, where in-game death is merely a near inconsequential and minor setback (Allison et al., 2015; Copcic et al., 2013). Some theoretical works have discussed the appeal of highly unfair and punishing games: Lazzaro (2004) argued that effort and frustration are needed to feel a personal triumph over adversity by overcoming difficult obstacles – the so-called fiero state, which leads to hard fun. Another study stated that unfair games can be funny because they are user-unfriendly and break every good-practice-level design rule (Wilson & Sicart, 2010). Furthermore, other studies have been devoted to the permadeath mechanism (Allison et al., 2015; Carter et al., 2013; Copcic et al., 2013) and depict a similarly punishing death mechanism as *Dark Souls*; however, the lack of permanence and the overall high difficulty gameplay in *Dark Souls* make it substantially different from games such as *DayZ*.

While these studies provide some valuable insight, their findings come mainly from theoretical discussions and empirical evidence is still scarce. Hence, the paucity of theoretical knowledge provided by the existing literature required a mixed-method approach: The roles of avatar death and high challenges in regard to positive experiences had yet to be empirically and quantitatively explored with the help of open questions, as they seem to play a key role in creating enjoyment through meaningful learning. The exact methods, results and discussion will now be presented in the summary of the first manuscript.

Summary of Manuscript 1: A Good Reason to Die: How Avatar Death and High Challenges Enable Positive Experiences

Aim of the study and contribution. Although the appeal of difficult or punishing games has received some attention in current research (Allison et al., 2015; Carter et al., 2013; Lazzaro, 2004; Wilson & Sicart, 2010), empirical evidence is still scarce. Given the crucial role of challenge-skill balance for positive experiences and the fact that the literature points mainly in a direction where excessive challenges are detrimental to the players' enjoyment (Gutwin et al., 2016; Klarkowski et al., 2016; Schmierbach et al., 2014), the present study aimed for a better understanding of the roles of avatar death and high challenges in regard to positive player experiences. As the majority of preexisting work plead for balanced experiences, this study aimed to explore the seemingly contradictory situation of players enjoying a game defined

by high challenges, numerous frustrations, and punishing avatar death. The central question is, do players enjoy high-challenge games despite the difficulties and failures or do avatar death and high challenges actually form and enable positive experiences? With the help of a mixed-method design, the present study was able to describe the roles of avatar death and high challenges by identifying their connection to important predictors for positive experiences, and therefore to put the results of previous studies conducted by Schmierbach et al. (2014), Klarkowski et al. (2016) and Gutwin et al. (2016) into perspective, which simply showed lower enjoyment scores for higher difficulties. In addition, the present study was able to describe the death mechanism as a learning process not only using qualitative reports as did Allison et al. (2015) and Carter et al. (2013), but also by further analyzing the frequencies of important themes connected to it.

Methods. The methods were accordingly chosen to gather the data needed to answer the questions resulting from the research gap identified, thus resulting in a mixed-method approach. To improve our understanding of why players enjoy games they constantly fail at and struggle with, we aimed for a very specific sample comprising fans of the game *Dark Souls III*, a game which is notorious for its high difficulty and punishing avatar death. A total of 95 participants were recruited from various social networks (e.g. *Facebook*, *Twitter*, *Vkontakte*) and gaming-related groups (e.g. *Reddit*, *Facebook*) and were asked to complete an online survey consisting of both qualitative open-ended questions and quantitative scales.

Qualitative open-ended questions. The lack of empirical studies for highly challenging gameplay and the role of avatar death suggested an explorative procedure consisting of collecting qualitative data in order to identify further important but yet unknown dimensions of interest. Therefore, the critical incident method (Flanagan, 1954) was applied. This allows for collecting qualitative data from open questions, evaluating important categories using the thematic analysis protocol (Braun & Clarke, 2006) and further analyzing these categories quantitatively. The open-ended questions followed a similar approach to that of Bopp et al. (2016). We asked participants to bring to mind an outstanding positive or negative experience they had had in a recent game session in *Dark Souls III*. Additionally, in a follow-up question they were asked to try to describe this particular experience as accurately, in as much detail and as concretely as possible in at least 50 words and to clarify the reason for these thoughts and feelings. Some categories for the content analysis emerged from a deductive theoretical standpoint and the consideration of

background literature, while other categories were developed inductively by exploring the open answers and reviewing them for important content dimensions. The open-ended answers were manually coded following the thematic analysis protocol (Braun & Clarke, 2006). The most common themes identified were *achievements & victories*, *learning & improvement*, *difficulties & failures*, *lack of progress* and *enemy encounters*. To assure interrater reliability, an independent rater coded a random subset of 41 experiences. The category *lack of progress* was dropped due to low agreement among the raters but subsequently substantial agreement among the raters for all themes ($k = .6$) was achieved. Furthermore, the overall valence of the experience was coded either positive or negative, depending on whether the play session was mainly described with positive outcomes such as joy, satisfaction, happiness and positivity or negative outcomes such as frustration, anger, anxiety or sadness. The interrater agreement for overall valence was also found to be substantial ($k = .65$).

Quantitative standardized scales. The following 7-point Likert-type scales, ranging from strongly disagree (1) to strongly agree (7), were included in the online survey mainly to acquire descriptive knowledge of our sample: positive and negative affect (I-PANAS-SF; Bateman et al. (2011)), Player Experience Need Satisfaction (PENS, Ryan et al. (2006)), challenge (Game Experience Questionnaire, IJsselsteijn et al. (2008)), challenge-skill balance (Schmierbach et al., 2014), and enjoyment (Oliver & Bartsch, 2010).

Results. The identified themes *achievements & victories*, *learning & improvement*, *difficulties & failures*, *lack of progress* and *enemy encounters* made up 87.8% of all reports, thus covering the most substantial part of all experiences. Out of all 95 experiences, 57 (60%) were coded as overall positive and 38 (40%) as overall negative experiences. These two groups, split by valence, were used in the subsequent analysis to investigate the characteristics of positive and negative experiences. The contingencies in Table 1 depict the frequencies of each theme split by overall valence.

Analysis of identified themes. The theme *achievements & victories* was coded in 61% of all experiences. These moments were usually reported after defeating certain enemies and rarely outside the context of a fight. The players' victories were usually described in contrast to the high challenges, the unpredictable outcomes, previous deaths and the consequences emerging from them. Another typical characteristic of this theme was its depiction in the light of fear and anxiety of lost

Table 1
The overall absolute and relative frequencies of the identified themes are depicted in the second column. In the third and fourth column these frequencies are split by valence of the experience (positive and negative) and tested for statistical significant differences with Pearson's chi-squared tests with Yates' continuity correction. Themes with statistical significant differences are shown in bold.

Theme	Overall (N=95)	Positive (n=57)	Negative (n=38)	χ^2	p
A&V	**58 (61%)**	**45 (79%)**	**13 (34%)**	**17.355**	**<.001**
L&I	**43 (45%)**	**37 (65%)**	**6 (16%)**	**20.268**	**<.001**
D&F	78 (82%)	49 (86%)	29 (76%)	0.863	.353
EE	64 (67%)	43 (75%)	21 (55%)	3.354	.067

Note. A&V = achievements & victories. L&I = learning & improvement. D&F = difficulties & failures. EE = enemy encounters.

progress. The theme *learning & improvement*, reported in 45% of all experiences, typically comprised moments of figuring out strategies, attack patterns and certain gameplay elements in order to progress within the game. These moments were typically evoked by avatar death and the high challenges the game imposes on the player. The most frequent theme *difficulties & failures* was reported by 82% of all participants. These moments typically described occurrences of avatar death, failed attempts to progress within the game, and the struggle and coping with high challenges as a result. This theme was usually reported together with *achievements & victories* and *learning & improvement*. The theme *enemy encounters* was reported in 67% of all experiences. It typically contained a narrative about a boss fight or a regular enemy in the game. Numerous reports also depicted a strong interrelationship among all themes when describing experiences, where all themes together played a substantial role.

The Pearson chi-squared test with Yates continuity correction (see Table 1) revealed significantly more observations of *achievements & victories* ($\chi^2 = 17.36$, $df = 1$, $p < .001$) and *learning & improvement* ($\chi^2 = 20.27$, $df = 1$, $p < .001$) in positive experiences compared to negative experiences, while *difficulties & failures* and *enemy encounters* did not differ significantly between overall positive and negative experiences. Furthermore, the theme *difficulties & failures* occurred significantly more often in experiences containing moments of *achievements & victories* compared to experiences without *achievements & victories* ($\chi^2 = 14.26$, $df = 1$, $p < .001$). Similarly, *difficulties & failures* occurred significantly more often in reports

of *learning & improvement* than in reports that did not include that theme ($\chi^2 = 7.8$, $df = 1$, $p < .01$).

Analysis of quantitative measures. The statistical comparisons of player experience measures split by valence (positive and negative experiences) were done by comparing the groups on an ordinal scale using Mann-Whitney U tests. The tests revealed that a positive experience is associated with greater positive affect ($Z = 4.30$, $p < .001$, $r = .44$), competence ($Z = 3.24$, $p < 0.01$, $r = .33$), relatedness ($Z = 1.95$, $p = .052$, $r = .20$), challenge-skill balance ($Z = 2.53$, $p < .05$, $r = .26$), challenge ($Z = 2.56$, $p < .05$, $r = .26$) and enjoyment ($Z = 3.75$, $p < .001$, $r = .38$). Although these measures were rated significantly higher in positive than in negative experiences, it is important to note that the ratings were usually also relatively high for negative experiences and that some of these differences, although significant, were rather marginal. See Table 2 for descriptive statistics.

Prediction of overall valence. To gain an understanding of the relative importance of all identified themes and quantitative measures, a binominal logistic regression was conducted to identify significant predictors of overall valence. The regression model revealed that only *achievements & victories* (Wald's $\chi^2 = 3.90$, $p < .05$, odds ratio = 3.9) and *learning & improvement* (Wald's $\chi^2 = 5.70$, $p < .05$, odds ratio = 4.7) were significant predictors. Accordingly, the occurrence of *achievements & victories* raised the chance of having a positive experience nearly four times. Similarly, *learning & improvement* led to an almost five times higher

Table 2
Mean, standard deviation, and median of player experience scales for experiences split by valence of the experience. Item sources: [1]I-PANAS-SF (Bateman et al., 2011), [2]PENS (Ryan et al., 2006), [3]GEQ (IJsselsteijn et al., 2008), [4]Schmierbach et al. (2014) and [5]Oliver and Bartsch (2010).

	Positive valence ($n = 57$)			Negative valence ($n = 38$)		
Scale	Mean	SD	Median	Mean	SD	Median
Positive affect[1]	5.93	0.75	6	4.88	1.28	4.8
Negative affect[1]	3.05	1.13	3	3.25	1.26	3
Competence[2]	5.79	0.77	6	4.87	1.41	5
Autonomy[2]	5.76	0.96	6	5.48	1.55	6
Relatedness[2]	4.45	1.33	4.67	3.81	1.58	6
Challenge[3]	6.17	0.76	6.2	5.38	1.46	6
Challenge-skill balance[4]	5.96	0.86	6	5.35	1.15	5.67
Enjoyment[5]	6.87	0.37	7	6.11	1.38	6.83

chance of having a positive player experience. Meanwhile, none of the quantitative measures had any significantly predictive value.

Discussion and conclusion. Even though punishing avatar death and high challenges are a substantial part of the game *Dark Souls III*, the play sessions were perceived as enjoyable, as reflected in the high ratings on enjoyment and positive affect. The vast majority of players in this study reported numerous negative events, however *difficulties & failures* did not occur significantly more often in either negative or positive player experiences, suggesting that failure or avatar death is not an exclusive characteristic of either of them. Moments of *achievements & victories* and *learning & improvement* occurred significantly more often in positive experiences though and they both were important predictors of positive reports.

The strongest predictor of a positive player experience was *learning & improvement*, supporting the notion that learning processes are critical to video game enjoyment (Flynn-Jones, 2015; Gee, 2005; Juul, 2013). Thus, moments of learning seem to have an especially important role in a game with very high difficulty. Furthermore, 95% of all participants who reported *learning & improvement* also reported *difficulties & failures*. As seen in the players' reports, learning processes largely stem from failures, which coincides with the statement from Juul (1999) that avatar death is the death you survive and learn from. Taken together these results emphasize the important role of high challenges and avatar death, which more likely cause players to fail, and therefore may enable learning processes which are critical to the players' enjoyment. The second predictor for positive experiences was *achievements & victories*. Similar to previous work on permadeath in *DayZ* (Allison et al., 2015; Carter et al., 2013), players in the present study often rated their achievement as particularly satisfying in view of previous failed attempts and struggle. Player reports showed that possible severe consequences resulting from high challenges and avatar death formed a general atmosphere of anxiety and fear of loss, which is in line with the general assumption that challenges beyond the challenge-skill balance lead to anxiety (Csikszentmihalyi, 1990). However, in this case anxiety emphasized achievements as hard-earned success. Similar to learning processes, 95% of participants who reported *achievements & victories* also reported *difficulties & failures*, which further showcases the interplay of positive and negative experiences, suggesting that not only are they not mutually exclusive, but that one actually to a large extent depends on the other. Our results indicate that difficult games are not in general less enjoyable, as assumed by Schmierbach et al. (2014). Not only

did *difficulties & failures* occur as often in positive as in negative experiences, but for many players they enabled and formed *achievements & victories* and moments of *learning & improvement* for positive experiences, thus demonstrating how closely negative and positive events are intertwined. In conclusion, *learning & improvement* and *achievements & victories* are the two most important predictors and thus directly linked to positive experiences; however, both are enabled and characterized by avatar death and high challenges. Whereas Bopp et al. (2016) showed that negative emotions such as sadness directly contribute to positive player experiences, the results of the present study in a similar vein suggest that negative events such as avatar death and high challenges do not directly predict a positive experience; rather they enable and characterize moments of achievement and learning.

Some important limitations have to be addressed though. To explore why some players enjoy video games with excessive difficulty we specifically recruited participants from fan forums of *Dark Souls*, which most likely led to a very specific and experienced sample with a strong positive bias towards this kind of game. The participants in this study most likely have a high tolerance for high difficulty gameplay. Hence, the results of this study may be dependent on personal preferences and cannot be generalized to all types of player. Moments of *achievements & victories* and *learning & improvement* may therefore only be crucial elements of positive experiences for so-called challenge seekers (Yun et al., 2010). Furthermore, following Juul's notion of players' repertoire (Juul, 2011), the perception of difficulty is shaped by the players' previous experiences, and the skills and strategies they have acquired. It would therefore be up to future research to explore how different player personalities with different skill repertoires influence the meaning of *achievements & victories* and *learning & improvement*. It would also be up to future work to compare different game difficulty levels or game genres with regard to the dimensions identified in order to prove causal effects.

Conclusion. The mixed-method approach revealed that high challenges and avatar death did not directly create positive experiences. However, for those players who reported a positive experience, they played a key role in forming *achievements & victories* and moments of *learning & improvement*, which in turn enabled a positive experience. Victories and learning were enjoyed so much because they had to be earned the hard way. These findings emphasize the roles played by *achievements & victories* and *learning & improvement* in a highly challenging context in order to attain positive experiences.

Data Quality from Crowdsourcing Platforms

The second manuscript revolves around warranting good data quality when recruiting participants from crowdsourcing platforms. Data collection in psychology and HCI-research is increasingly conducted using online surveys. In the years 2003 and 2004, just 1.6% of all studies published in APA journals made use of the internet (Skitka & Sargis, 2006), while roughly ten years later Gosling and Mason (2015) stated that it would be impossible to review all studies using the internet for data collection and that this method covers basically all areas of psychology. With the advent of crowdsourcing services such as *Amazon's Mechanical Turk* (*MTurk*) or *FigureEight* (formerly known as *CrowdFlower*), where various studies can be offered to so-called crowd-workers, online data collection has become more popular than ever (Kan & Drummey, 2018). Chandler and Shapiro (2016) estimated the number of published papers between 2006 and 2014, using crowdsourced online samples from *MTurk*, as being as high as 15,000.

Advantages and Disadvantages of Online Data Collection

The reasons for the increasing popularity of online data collection are evident when considering the advantages it holds in terms of its versatility over traditional data collecting methods such as face-to-face-interviews and pen-and-paper surveys conducted in the laboratory. First of all, online data collection is usually faster and cheaper (Casler et al., 2013; De Winter et al., 2015; Diekmann, 2009), mainly due to wider distribution of the study, lower hurdles for participation and lower infrastructure costs (Casler et al., 2013; De Winter et al., 2015; Kan & Drummey, 2018). Therefore, large sample sizes can be achieved with relatively low effort when compared to traditional recruitment methods. Aside from factors surrounding time and monetary resources, online data collection also holds some technical and organizational advantages over traditional methods. The inclusion of multimedia (e.g. pictures, videos, graphs, etc.) and filter questions is more convenient and randomized versions of a survey, used to conduct experiments or pretest a survey, are easier to apply in an online version (Diekmann, 2009).

These advantages of online data collection also apply to crowdsourcing services (Kan & Drummey, 2018): Casler et al. (2013) found that recruiting via social media networks or crowdsourcing platforms such as *Mechanical Turk* was faster than testing undergraduates and De Winter et al. (2015) came to the same conclusion, compar-

ing online data collection with recruiting via agencies. Furthermore, De Winter et al. (2015) compared the costs of online data collection with recruiting via a Dutch marketing agent and found that the agent's cost exceeded the costs of online data collection by more than 30 times. Moreover, crowdsourcing services offer a more diverse population compared to typically homogenous samples from psychological studies (Kan & Drummey, 2018). Crowd-workers from *MTurk*, for example, consist of a demographic containing more than 500,000 individuals from 190 countries (Paolacci & Chandler, 2014). Although concerns about the generalizability and validity of crowdsourced samples have also been discussed (Kan & Drummey, 2018), Gosling and Mason (2015) reported that the mean and range of age of a crowdsourced online sample on *Mechanical Turk* is more representative of the general US population than a sample merely consisting of undergraduate students. Other studies further reported that crowdsourced samples in particular have greater diversity and better balanced gender ratios when compared to online samples from social media (Casler et al., 2013; De Winter et al., 2015).

However, literature on data collection as well as numerous studies have also highlighted the downside of data collection with online surveys. Apart from the fact that online samples tend to be systematically distorted compared to the basic population (Diekmann, 2009), with such samples usually tending to be younger, overeducated, underemployed, less religious and more liberal in terms of their political views (Berinsky, Huber, & Lenz, 2012; Paolacci & Chandler, 2014; Shapiro, Chandler, & Mueller, 2013), increased attention from various researchers has recently been dedicated to inattentive and careless responding in online surveys (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). The increased distance between researchers and participants in a possibly distracting and uncontrolled environment most likely results in such samples containing deficient data quality, stemming from inattentiveness or other forms of deceptive behavior. Although this phenomenon may also occur in lab studies, crowdsourced samples seem to be especially prone to carelessness, as these respondents are usually non-naïve subjects who respond to studies in an uncontrolled and possibly distracting environment and the incentive structure of the platform tempts participants to engage in deceptive behavior (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Kan & Drummey, 2018; Peer et al., 2017; Stewart, Chandler, & Paolacci, 2017).

Careless Responding: Causes, Prevalence, and Effects

While the absence of a researcher may diminish unwanted interviewer and context effects on participants (Diekmann, 2009), the anonymity when participating in online studies also leads to a lack of accountability (Meade & Craig, 2012), which entails a whole new class of problems possibly resulting in insufficient data quality. However, this section and the related manuscript focus solely on inattentiveness or careless responding, which refers to participants not paying full attention or not putting enough effort in when reading instructions or answering questions. It is important to note, however, that there are also other forms of invalid responding such as social desirability, faking good, and faking bad, which also cause deficient data quality but may have different causes and effects on the study (Maniaci & Rogge, 2014; McKay et al., 2018). Apart from a lack of accountability, participants' interest, length of the survey, social contact, and environmental distractions are further causes of carelessness in online surveys (Meade & Craig, 2012). Furthermore, extrinsic motivation, stemming from monetary or other forms of incentives, result in participants minimizing their investment and effort when completing online studies (Gadiraju, Kawase, Dietze, & Demartini, 2015).

The estimated prevalence of careless responding varies between and within studies, depending on the measures used to detect carelessness and the origin of the online samples. A study conducted by Meade and Craig (2012), using an online sample consisting of students, estimated that roughly 10% to 12% of all participants in an online survey show careless answering behavior. Maniaci and Rogge (2014) found that 3% to 9% respond carelessly or inattentively in a mixed online sample. In an online survey with students from a university in the United States, Ward, Meade, Allred, Pappalardo, and Stoughton (2017) showed that 23% of the participants were flagged by at least one instructed response item. Dogan (2018) estimated a carelessness prevalence of 40.7% to 59.8% for a sample collected on *Facebook*. Regardless of these highly varying estimations of careless or inattentive behavior in online samples, even small amounts of careless responding can have grave consequences for a study, such as failed replications (Oppenheimer et al., 2009), identifying a nonexistent effect (Huang et al., 2015), failed manipulations when instructions are not carefully read (Maniaci & Rogge, 2014), lower internal consistency of validated scales (Maniaci & Rogge, 2014), and problems in developing a questionnaire and in item analysis (Johnson, 2005).

Most empirical research concerning careless and inattentive responding in the past years has given attention to the detection of this phenomenon (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012). The methods to detect participants with careless behavior can be divided into two groups: planned detection methods, such as attention checks and self-reported data quality, and post hoc measures, such as the longstring index or odd-even consistency (see Curran (2016) for an overview). The detailed measures applied in the second manuscript will be presented in the corresponding summary.

Research Gap: Prevalence and Task-Dependence of Carelessness in Crowdsourced Samples

Although online data collection has become a standard procedure over the past few years, research on methods ensuring data quality is still sparse. Various studies concerning the detection of careless respondents have analyzed different measures and provided concrete recommendations (Maniaci & Rogge, 2014; Meade & Craig, 2012; Niessen et al., 2016), however, many questions still remain open. Firstly, most recent research has examined academic participant pools or mixed types of online data (Maniaci & Rogge, 2014; Meade & Craig, 2012) and estimations of careless and inattentive responding in crowdsourced samples remain largely unknown. Studies analyzing data quality on crowdsourcing platforms have assessed other forms of deceptive behavior, such as faking eligibility requirements (Kan & Drummey, 2018), or they have merely applied one measure to detect carelessness (Hauser & Schwarz, 2016; Peer et al., 2017), resulting in highly volatile estimations of inattentive behavior, ranging from 4% to 74.5%. Inattention in the study conducted by Hauser and Schwarz (2016) was assessed by merely applying an instructional manipulation check (IMC), a method which has been criticized for being too restrictive (Maniaci & Rogge, 2014), and Peer et al. (2017) analyzed carelessness in crowdsourced samples using only attention checks. While these studies provide some valuable insight into careless behavior on crowdsourcing platforms, they lack a variety of methods to analyze participants' inattention. Various other carelessness measures therefore have yet to be systematically assessed and discussed in the context of crowdsourced online samples.

Another open question revolves around the task dependence of careless behavior. In HCI-research and psychology, the inclusion of open-ended questions is a common way to assess experiences, attitudes or values towards a certain object or

topic by applying, for example, the critical incident method (Flanagan, 1954). This method is used to capture qualitative data and provide different perspectives on a phenomenon, as well as first insight into relatively new and undiscovered fields, as demonstrated in the first manuscript of this dissertation. Some studies showed that open questions produce more invalid answers and nonresponses than closed-ended measures (Allison et al., 2002; Holland & Christian, 2009; Reja et al., 2003) and nonresponses are usually associated with participants who were less interested in the topic of the question (Groves, Presser, & Dipko, 2004; Holland & Christian, 2009). However, it is still unknown how the quality of the open answers given relates to other measures of carelessness.

Accordingly, there is a lack of a systematic analysis of careless behavior on crowdsourced platforms, taking various measures to detect careless and inattentive behavior into consideration. The following section presents the methods, results, and discussion in regard to this challenge in order to make recommendations for future research.

Summary of Manuscript 2: Almost Half of the Participants in Online Surveys are Inattentive: An Investigation of Data Quality in Crowdsourced Samples

Aim of the study and contribution. The study of the second manuscript aimed for a better understanding of carelessness and inattention on crowdsourcing platforms and the task dependence of this phenomenon, addressing the limited variety of methods used in preexisting works and thus applying various planned detection methods and post hoc measures. The results of a latent profile analysis revealed that 45.9% of the sample could be classified as careless. This is an alarmingly high portion of the sample, as the exclusion of this many participants would raise methodological and economical concerns. Correlations between open answer quality and other carelessness measures proved to be rather low, demonstrating that careless behavior would seem to be task dependent and, thus, task-related measures to detect inattentive behavior also have to be taken into consideration when analyzing the data quality of a sample. The second manuscript provides subsequent recommendations for future research conducted with crowdsourced samples: The conditional inference tree analysis revealed the instructed response item (IRI), bogus item, and the open answer quality assessment to be important and precise predictors for detecting participants in the careless class from the latent profile analysis.

Methods. The study was conducted by recruiting a sample of 394 participants on the crowdsourcing platform *FigureEight*. The online survey started with an open-question task about a recent negative online shopping experience, where participants were asked to answer in as much detail as possible, using complete sentences and at least 50 words. After a set of various scales, namely, the PANAS (Watson, Clark, & Tellegen, 1988), AttrakDiff2 (Hassenzahl, Burmester, & Koller, 2003), and psychological need satisfaction (Sheldon, Elliot, Kim, & Kasser, 2001), either a high trust or a low trust mockup of a website was presented and manipulated according to trust elements identified by Seckler et al. (2015). The mockup websites subsequently had to be rated with scales concerning trust (Casalo, Flavián, & Guinalíu, 2007) and visual aesthetics (VisAwi, Moshagen and Thielsch (2010)), followed by the Big Five Inventory (BFI) (John, Donahue, & Kentle, 1991), demographic information, and various scales and items for self-reported carelessness (Maniaci & Rogge, 2014; Meade & Craig, 2012).

Planned detection methods and post hoc measures. The measures used in this study for detecting careless and inattentive behavior can be divided into two groups (see Curran (2016) for an overview). The first group comprises planned detection methods in terms of which special items or scales are implemented that help to identify careless responding: attention check questions such as the bogus item and the IRI (Curran, 2016; Meade & Craig, 2012) as well as questions assessing self-reported data quality: self-reported careless responding, patterned responding, rushed responding, skipping of instructions, and self-indicated data usage with a self-reported single item (SRSI UseMe) (Meade & Craig, 2012). The second group of methods detecting careless responding comprises post hoc measures which do not require the implementation of special items. This group includes measuring the overall response time (Curran, 2016; Huang et al., 2015; Maniaci & Rogge, 2014), measuring strings of identical answers with the longstring index (Curran, 2016; Huang et al., 2015), assessing answer inconsistencies with the odd-even consistency (OEC), resampled individual reliability (RIR), and the correlation of a person's answer with the mean of answers given by the whole sample, as shown by the person total correlation (PTC) (Curran, 2016).

Open answer quality. The open answer quality was rated by applying indicators from Holland and Christian (2009) and Smyth, Dillman, Christian, and McBride (2009). The indicators assessed 1. whether participants provided a thematically substantive answer, 2. whether they provided at least 50 words, 3. whether

they provided answers in complete sentences, 4. the number of subquestions answered, and 5. the number of subquestions elaborated. All indicators were calculated into an Open Answer Quality Index, which showed excellent inter-rater agreement ($ICC3$ = .96, with a 95% confidence interval from 94. To .97 (F(99,99)=51, $p <$ .001). A detailed description of building the Open Answer Quality Index is provided in the appendix section of the second manuscript.

Results. ***Planned detection methods and post hoc measures.*** Recommended cutoffs by Maniaci and Rogge (2014) or Curran (2016) were used to flag participants with each measure. Thus, the self-reported items revealed that 25 (6.6%) participants were flagged for carelessness, 50 (12.7%) were flagged for patterned responding, 44 (11.2%) for rushed responding and 65 (16.5%) for skipping instructions. The SRSI UseMe item was negated by 22 (5.6%) participants. An aggregated self-report measure revealed that 106 participants (26.9%) indicated self-reported carelessness. The IRI and bogus item were missed by 96 (24.4%) and 92 (23.3%), respectively. Both attention checks together were missed by 52 (13.2%) participants. Looking at post hoc measures, no suspiciously fast respondents were flagged, the recommended cutoff for the longstring analysis though flagged 22 (6.3%) participants. Similarly, 63 (16%) participants failed to meet the OEC cutoff and 61 (15.5%) were revealed as answering too inconsistently in the RIR. Lastly, the PTC flagged 74 (18.8%) as careless respondents. See Table 5 for an overview.

Open answer quality and its relation to other measures. Open answer quality was either coded 0 (=insufficient), 1 (=high), or 2 (=excellent). 100 participants (25.4%) provided insufficient answer quality in the open question. These participants failed the IRI ($\chi^2(df = 1, \text{N} = 394) = 21.35$, $p = 3.826\text{e}{-06}$) as well as the bogus item ($\chi^2(df = 1, \text{N} = 394) = 24.665$, $p = 6.82\text{e}{-07}$) significantly more often in comparison to participants providing high or excellent open answer quality. Moreover, considerably more participants from this group (43%) were flagged by self-reported carelessness compared to the other two groups combined (21.4 %). A Wilcoxon rank-sum test yielded significantly higher longstring values at an alpha level of 5% ($W = 17730$, $p < 0.01$) for the group providing insufficient open answer quality. Similarly, participants failing to provide sufficient answer quality failed considerably more often to meet the OEC, RIR, and PTC cutoff when compared to the rest of the sample. Lastly, these participants also answered in significantly ($W = 10958$, $p < 0.01$) less time ($M = 882.58$ seconds, $SD = 563.66$ seconds, n = 97) compared to the subsample providing sufficient answer quality ($M = 1042.83$, SD

= 544.93, n = 289). OEC showed relatively low correlations with other measures of carelessness, ranging from .13 (OEC) to .26 (bogus item, IRI+bogus item).

Latent profile analysis and class prediction. Following the approach as described and used by Meade and Craig (2012) and Maniaci and Rogge (2014), a latent profile analysis was conducted to identify different classes of carelessness in the sample. The latent profile analysis revealed three classes: class 1 (n = 181, 45.9%), class 2 (n = 129, 32.7%), and class 3 (n = 84, 21.3%). As seen in Table 3, class 1 is clearly associated with careless responding, as this class exclusively missed both attention check items (IRI, bogus item), self-reported carelessness more frequently, and is characterized by a very large longstring index and very low PTC. However, neither consistency items (OEC, RIR) nor response time offered a clear interpretation, as they seem to be associated both with class 1 and other classes.

A conditional inference tree analysis revealed open answer quality, the IRI and

Table 3
Descriptive statistics for each identified class of participants. IRI = Instruced Response Item, OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation.

Variable	Class 1	Class 2	Class 3
Class size	181 (45.9%)	129 (32.7%)	84 (21.3%)
Percentages pass			
Answer quality (%)	44.75	100	100
Bogus Item (%)	49.17	100	100
IRI (%)	46.96	100	100
Self-report (%)	59.12	90.70	76.19
SRSI UseMe (%)	90.06	99.22	96.43
Means			
Response time (in Minutes)	14.58	16.94	22.03
OEC	.52	.86	.37
RIR	.44	.83	.43
PTC	.13	.59	.30
LongString	9.61	3.79	4.83
Means (Self-reported)			
Careless responding	2.16	1.36	1.52
Patterned responding	2.28	1.20	1.49
Rushed responding	2.43	1.68	1.88
Skipping instructions	2.53	1.99	2.13

Note. Total $N = 394$

the bogus item to be precise predictors for separating class 1 from class 2 and class 3 (see Table 4). Applying this combination of measures as well as the SRSI UseMe and longstring index to flag participants led to a sample (n = 209, 53.05% of the total sample), which showed an increased effect size in the mockup website experiment and yielded a smaller p-value for an independent samples t-test, when compared to the full sample.

Discussion and conclusion. Almost 60% of all participants were flagged by at least one of the methods examined in this study (see Table 5). The planned detection methods and post hoc measures tended to flag a considerably higher percentage of participants in the crowdsourced sample as careless when compared to a previous study, applying these measures to a mixed online sample (Maniaci & Rogge, 2014). Roughly one quarter of all participants (24.4%) failed to answer the IRI correctly and 16% failed the OEC cutoff. These results clearly surpass the percentages of flagged participants in the study conducted by Maniaci and Rogge (2014), where 14% failed the IRI and 7% were flagged by the OEC measure. The longstring index revealed comparable results, flagging 6.3% in the fully crowdsourced sample from the present study and 6% in the study from Maniaci and Rogge (2014).

Thus, just looking at some measures individually might indicate that inattentive behavior in a fully crowdsourced sample may be more pronounced than in other types of online collected data. Moreover, the relative number of participants flagged with attention check items can be expected to increase with the length of the survey, as one attention check is recommended for every 50 to 100 items (Meade & Craig, 2012). In a considerably longer study, the application of four attention check items in a crowdsourced sample, Peer et al. (2017) resulted in 73% of all participants failing at least one of them. However, a method of combining all of these measures was needed in order to draw further conclusions.

Table 4
Performance of the conditional inference tree model in predicting class membership.

	Class 1	Class 2	Class 3
Predicted			
Class 1	180	0	0
Class 2	0	126	1
Class 3	1	3	83

Note. Total $N = 394$

Hence, the latent profile analysis revealed one big class (class 1), containing 181 (45.9%) participants, that displayed careless behavior, which is well above the estimated 2.2% to 11% from the student pools and mixed online samples from Meade and Craig (2012) and Maniaci and Rogge (2014). Class 1 cannot be described by one measure alone, as it contains multiple forms of inattentive behavior. The careless class is, however, characterized by exclusively failing to provide sufficient open answer quality, as well as passing the IRI and bogus item. Class 1 self-indicated carelessness considerably more often, showed very large longstring values and a very low PTC, thus indicating excessive consistency within, yet low congruence with the total sample. Class 3 though, usually inconspicuous concerning other measures of carelessness, showed comparably poor RIR and even worse OEC than class 1. This finding suggests using measures of consistency as a means of data cleaning with caution, as they might have potential for a high false-positive rate.

Although good open answer quality was associated with significantly better results for self-reported data quality, attention checks, longstring index, OEC, RIR, and PTC, the correlations with these measures of carelessness were rather weak. This might point to the fact that carelessness does indeed depend to a large extent on the task given. This result coincided with findings from Maniaci and Rogge (2014), showing that carelessness during specific tasks, such as watching a video or marking pronouns in a text, mostly correlates low with other measures of carelessness.

Table 5
Descriptive statistics for all detection methods used in the study. Self-report includes problematic responding tendencies as well as the SRSI UseMe item.

	Mean	SD	Min	Max	No. Flagged	%
Planned detection						
Self-report					106	26.90
Bogus Item					92	23.35
Instructed Response Item					96	24.37
Response time	16.71	9.22	3.93	61.15		
Post hoc detection						
LongString	6.63	9.15	0	44	25	6.35
Odd-even consistency	.61	.43	-1	1	63	15.99
Resampled individual reliability	.56	.39	-.82	.99	63	15.99
Person-total correlation	.38	.32	-.47	.88	74	18.78
Answer quality					100	25.38
Total (flagged by ≥1 method)					233	59.14

Note. Total $N = 394$

Based on the latent profile analysis as a frame of reference for careless responding, the conditional inference tree for predicting class 1, and a pragmatic point of view, the study recommends the application of the SRSI UseMe, IRI, bogus item, longstring index and a task-specific measure to assess data quality in a crowdsourced sample. These measures either represented important and precise predictors for the inattentive class 1, or they provided pragmatic merit for analyzing the data quality of a crowdsourced sample: All these methods are relatively straightforward to apply, as they do not need to consider scale dimensions and inverse items. They accurately predicted the class 1 membership of 180 out of 181 participants, while none of the participants from class 2 and 3 were falsely flagged by this combination of methods. All other post hoc measures are not recommended, however, as they were not clearly associated with one class and not predictive of class 1 membership.

Still, some limitations have to be considered concerning these results and recommendations: Future research should systematically analyze carelessness on different crowdsourcing platforms and assess whether these results also apply to different services. Furthermore, as the exclusion of nearly half of the sample is neither methodologically nor economically reasonable, more research has to be done with the aim of preventing carelessness. It is also important to note that the study merely assessed careless behavior as other forms of invalid responding (e.g. faking good or faking bad) cannot be detected using these methods. Furthermore, the recommendations were largely based on class 1 as revealed by the latent profile analysis, including an open-question task. Whether this analysis would yield similar results with different tasks is also up to future research. Lastly, as all post hoc detection methods are approximate and uncertain, bad data quality can not clearly and reliably be identified in every case. Our recommendations are based on the prediction of class 1 membership in a latent profile analysis. Only planned detection methods were found to be predictive. However, there are situations where it might not be possible to include attention check items or task-dependent measures of quality, such as voluntary surveys or highly specific populations.

Conclusion. The aim of this study was to give an estimate of the prevalence of carelessness in samples from crowdsourcing platforms, based on different identification methods, and to make subsequent recommendations. Our results reveal that roughly half of all crowdsourced participants show careless behavior. Furthermore, carelessness and inattentive behavior seem to be highly task dependent, as correlations between open answer quality and other measures are rather low. Fi-

nally, based on a predictive model and interpretative problems of several measures, we recommend assessing the data quality of crowdsourced samples by applying an SRSI UseMe item, attention checks such as the IRI and bogus item, as well as the longstring index and a task-specific measure. The combination of these measures was able to precisely identify inattentive participants and the subsequent exclusion of this subsample led to increased effect size and smaller p-values in an experiment.

Measuring Trust on the Web

The third and last manuscript of this dissertation depicts the development and validation of a differential for measuring trust on the web. Trust in research has been discussed across a wide range of disciplines, which has led to an overabundance of definitions from various academic fields (Van der Werff et al., 2018). This makes a precise operationalization for measuring trust rather difficult. In psychology it is generally agreed that trust is an important concept and vital to personality development (Erikson, 1963), cooperation (Deutsch, 1962), and social life (Rotter, 1967). The majority of all trust definitions usually share two key components: that trustworthiness includes a willingness to be vulnerable and a perception of the intentions of the other party (Lewicki & Brinsfield, 2012). A ubiquitous definition of trust is proposed by Rousseau, Sitkin, Burt, and Camerer (1998) who describe trust as "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another". Furthermore, trust is an essential factor, allowing people to act under conditions of uncertainty and with the risk of negative consequences (Casalo et al., 2007).

Characteristics and Dimensions of Trust on the Web

The characteristics and dimensions describing trust in the context of the web primarily refer to a buyer-seller relationship in e-commerce. There are four characteristics of trust on the web, which are generally observed and accepted (Wang & Emurian, 2005). First, there must be two specific parties in a trusting relationship – a trusting party (trustor, e.g. an online customer) and a party to be trusted (trustee, e.g. an online merchant). Second, trust involves vulnerability, uncertainty and risk on the trustor's side, while anonymity and unpredictability are associated with the trustee. Third, trust leads to actions which are comprised mainly of risk-taking behaviors such as providing personal and financial information. Finally, trust is a subjective matter; the level of trust considered sufficient to make transactions online is different for everyone and people hold different attitudes toward machines and technology (Wang & Emurian, 2005). Additionally, users' trust on the web comprises three core dimensions – benevolence, integrity, and competence – which are described in the Web Trust Model developed by McKnight et al. (2002b) and which are generally agreed on (Bhattacherjee, 2002; S. C. Chen & Dhillon, 2003; Flavián et al., 2006; Gefen, 2002; Mayer, Davis, & Schoorman, 1995; McKnight et

al., 2002a). Benevolence is related to the user's belief that the other party is interested in his welfare, motivated by a search for a mutually beneficial relationship and without the intention of opportunistic behavior. In other words, the website owner is concerned with the present and future interests, desires and needs of their users and gives useful advice and recommendations. Integrity, sometimes referred as honesty, is the belief that the other party will keep his or her word, fulfill promises, and be sincere. This means, that no false statements will be made and all information you receive is honest. Competence on the other hand, sometimes referred to as ability, means that the website owner has the resources and capabilities needed for the successful completion of a transaction and the continuance of the relationship (Casalo et al., 2007). These core dimensions help an individual to collect evidence to make a trust decision and eventually engage in trust behavior (McEvily, Perrone, & Zaheer, 2003). Some literature also mentions either predictability or value congruence as possible fourth subdimensions (Dietz & Den Hartog, 2006; McKnight et al., 1998), although others argue that parts of these concepts are already included in integrity (Van der Werff et al., 2018).

Differences Between Online and Offline Trust Relationships

While online and offline trust relationships resemble each other, both involving conditions of risk and vulnerability, they also differ in various respects. Online trust not only usually comprises an interpersonal trust relationship but is additionally complicated by trust in the internet technology itself and the organization behind the technology (Van der Werff et al., 2018). Thus, the main differences between online and offline trust can be explained by discussing the trust relationship between these three agents (individuals, organizations, and technology). Furthermore, online trust is largely characterized by a lack of face-to-face interaction and an information asymmetry between the consumer and seller (Van der Werff et al., 2018). The physical separation between the trustor and the trustee prevents parties from monitoring previous behaviors and gaining substantial information for making a trust decision, leaving buyers with high levels of uncertainty (Pavlou, Liang, & Xue, 2007). Additionally, trust in the web is characterized by heightened privacy concerns and vulnerability, which relate to the fear that other individuals or malicious third parties will misuse previously disclosed information such as private communications or customer data (Van der Werff et al., 2018). Thus, activities like online shopping are perceived to involve more risk than traditional shopping (Lee & Turban, 2001).

Interpersonal Trust, Organizational Trust, and Trust in Technology

Interpersonal trust, organizational trust and trust in technology consist of characteristics that are exclusive to trust on the web. While interpersonal trust may function similarly online as in an offline context, some differences have to be considered. Depending on different modes of online communication (e.g. video call, live chat or e-mailing), temporal, spatial and geographic separation represents a possible hurdle for building trust (Van der Werff et al., 2018), which also increases the likelihood of deceptive behavior (Naquin, Kurtzberg, & Belkin, 2010). Research in this domain is mainly about identifying factors that lower trust barriers, e.g. applying avatars for online communication (e.g., Bente, Rüggenberg, Krämer, & Eschenburg, 2008).

Trust in online organizations is formed by a website's structure design, graphic design, content design and social cue design (Wang & Emurian, 2005). Examples of such factors influencing users' trust are color tones and color balance (J. Kim & Moon, 1998), social presence cues such as human photos or personalized greetings (Hassanein & Head, 2007), display of a strong privacy policy (Lauer & Deng, 2007), as well as the overall structure and accessibility of information (Bart et al., 2005; Flavián et al., 2006; Koufaris & Hampton-Sosa, 2004), to name just a few. Seckler et al. (2015) argue that while website design influences perceptions of distrust, trust is largely based on social determinants such as friends' recommendations or reviews.

Trust in technology systems has yet to be further explored as it still is a lesser understood area of research, which primarily has been conducted in computer sciences and information systems research (Van der Werff et al., 2018). Trust in technology revolves mainly around faith in its reliability, functionality and helpfulness and the belief in positive outcomes which influence online behavior (McKnight et al., 2011). The distinction between interpersonal trust and trust in a technology may not always be as clear though, for instance Wang and Emurian (2005) and Lankton and McKnight (2011) found that consumers treat technological agents, such as recommendation agents or a social network like *Facebook*, as social actors with whom they form social relationships.

Whether and to what degree the three trust dimensions, benevolence, integrity, and competence, are applicable to all three of these agents in the internet context (individual, organization and technology) is still a subject of debate, as a technology system is not a conscious trustee in which to build trust (Friedman, Khan Jr,

& Howe, 2000). Some proposed models for trust in IT systems thus include subdimensions like performance, helpfulness, predictability, or functionality (McKnight et al., 2011; Söllner et al., 2013), which address the fact that trust in technology systems is largely determined by functional, technical dimensions and less by value congruence and interpersonal expectations. Taken together, it comes as no surprise that a theoretical background with this level of versatility and complexity spawned an overabundance of scales to measure trust on the web, which leads to certain methodological issues.

Existing Scales for Measuring Trust in the Web Context

The variety of academic fields and contexts depicted above, where trust has been a subject of discussion, has led to a multitude of questionnaires and scales measuring trust (e.g., Bart et al., 2005; Bhattacherjee, 2002; Cho, 2006; Corbitt et al., 2003; Flavián et al., 2006; Lee & Turban, 2001; McKnight et al., 2002a; Pavlou & Gefen, 2004). However, one of the main issues of online trust research is the lack of a common, validated, reliable, versatile, and easy-to-translate measure including the three dimensions of benevolence, integrity, and competence (Y. Kim & Peterson, 2017). The problems associated with these preexisting scales can be summarized as follows: First, not all of these questionnaires directly assess trust but rather ask about adjacent constructs such as benevolence (Cho, 2006), which according to McKnight et al. (2002b) is merely one subdimension of the model for trust. Second, the applicability of these measurements in research and practice is limited since these scales were created to assess trust in a very specific context. For instance, the items developed by McKnight et al. (2002a) are tailored to a specific website under examination (e.g., "LegalAdvice.com is competent and effective in providing legal advice."). Another example is from Lu, Wang, and Hayes (2012), who developed Likert-type questions for customer-to-customer (C2C) platforms, such as: "Do you agree that this C2C platform solves a security problem or stops a fraudulent behavior." Applying such items in a different context requires extensive rephrasing, which in turn may be accompanied by a loss of reliability and validity. Furthermore, translating Likert-type items into other languages can be a difficult and time-consuming process which may further affect validity. Third, Y. Kim and Peterson (2017) stated that preexisting measurements were "ambiguous" and that there is a necessity for a "well-developed scale to measure online trust that is specifically tailored to the business-to-consumer e-commerce environment".

Research Gap: A Validated and Easy-To-Apply Semantic Differential for Trust on the Web

The above-mentioned problems with preexisting scales can be addressed with the development of a semantic differential for measuring trust on the web. Semantic differentials consist of multiple sets of bipolar items, usually a pair of adjective antonyms, which gives them several advantages over traditional Likert-type scales (Verhagen et al., 2015). Respondents to Likert-type scales can only indicate the extent to which they agree or disagree with a specific statement, thus agreeing with an item does not necessarily imply that the respondent disagrees with the opposite of this item (Chin, Johnson, & Schwarz, 2008). The range of possible answers in semantic differentials is more exhaustive and therefore allows the respondent to express any opinion without a suggested item formulation. Furthermore, semantic differentials can reduce the acquiescence bias provoked by Likert-type scales, a category of response biases indicating that respondents have a tendency to agree with all items or to indicate a positive connotation (Friborg, Martinussen, & Rosenvinge, 2006). Semantic differentials additionally outperform Likert-type scales on robustness (Hawkins, Albaum, & Best, 1974), reliability (Wirtz & Lee, 2003), and validity (Van Auken & Barry, 1995). And lastly, semantic differentials consist merely of two bipolar words per item, which reduces survey completion time (Chin et al., 2008) and suggests a format which is easier to apply in different contexts and easier to translate into different languages. Taken together, semantic differentials represent a promising tool for assessing trust in various domains of research and practice, which led to the present study developing the *TrustDiff* – an easy-to-apply measurement of trust including the three dimensions of benevolence, integrity, and competence. The exact methods used for the development and validation of the *TrustDiff*, as well as the results and discussion, will be presented in the following section.

Summary of Manuscript 3: TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web

Aim of the study and contribution. The manuscript includes three independent studies describing the development and validation of the *TrustDiff* – a semantic differential for measuring trust on the web. As depicted above, trust has been a subject of research in various academic fields, which has led to a multitude of definitions, proposed dimensions and measuring scales, even just within the domain

of trust in the web context. There is nevertheless a lack of a common validated and reliable measure including the three dimensions of benevolence, integrity, and competence (Y. Kim & Peterson, 2017), which mainly stems from preexisting measures being context-specific Likert-type scales relating to different dimensions (e.g., Lu et al., 2012; McKnight et al., 2002a). The results of three validation studies indicate that the *TrustDiff* has excellent psychometric properties, measuring benevolence, integrity, and competence with high reliability and good structural validity. The manuscript further demonstrates how the *TrustDiff* is related to an existing Likert-type trust scale and the concepts of visual aesthetics and usability. The *TrustDiff* was also found to be sensitive to the manipulation of trust-related features in an experiment with a mock website, showing significant differences on all subscales, thus demonstrating criterion validity. Finally, the *TrustDiff* as a semantic differential promises to hold several advantages over traditional Likert-type scales (Verhagen et al., 2015), as it is easier to translate into different contexts and languages. Accordingly, the *TrustDiff* represents a promising tool for assessing trust in various domains of research and practice.

Methods. The development and validation of the *TrustDiff* differential was conducted in three independent studies, following the framework described by Verhagen et al. (2015). To create an initial item pool, an extensive literature and scale review was conducted, extracting key adjectives from preexisting questionnaires. Possible antonyms were then selected, and near duplicates were removed, resulting in 28 positive adjectives with up to three different antonyms. An expert panel consisting of psychologists and UX researchers (N = 18) subsequently conducted an item-sort task for the three dimensions of benevolence, integrity, and competence, excluding adjectives with less than 13 correct assignments (Howard, 2016). For each of the remaining adjectives, the best fitting antonym with the highest agreement was chosen, resulting in an initial item pool of 20 items.

Study 1: Factor analysis, convergent validity, and discriminant validity. The goal of study 1 was to employ a factor analysis in order to reduce the over-representative item pool as well as to test the convergent and discriminant validity of the scale. A total of 601 participants were recruited from *Mechanical Turk* to conduct an online survey. The participants were asked to perform two information search tasks on one of two randomly assigned and unknown websites in order to prevent any biases from previous experiences. Upon returning to the survey, participants were asked to rate the website on the 20 items of the *TrustDiff*. To further

assess the convergent validity, a Likert-type trust scale (Flavián et al., 2006) was included in the survey. To assess the discriminant validity, participants were also asked to rate the website's usability (UMUX, Borsci, Federici, Bacci, Gnaldi, and Bartolucci (2015)) and visual aesthetics (VisAWI, Moshagen and Thielsch (2010)).

Study 2: Confirmatory factor analysis. The second online study with 312 participants from *Mechanical Turk* was conducted to test the measurement model with a confirmatory factor analysis. Participants were asked to name a single interactive technology they use frequently and to indicate how often they had used this particular technology over the last 14 days. The 14 remaining *TrustDiff* items from study 1 were included to rate this technology.

Study 3: Criterion validity. The goal of the third study was to test criterion validity by manipulating a website regarding its trust features and assess whether the *TrustDiff* is able to differentiate between those two versions. A total of 252 participants were recruited on *Figureeight* and asked to rate one of two randomly assigned versions of a mock online shop based on a screenshot provided. One version of the website included several trust-supporting elements while the second version lacked of any trust-supporting cues. The manipulation of the website characteristics was conducted according to Seckler et al. (2015) and Wang and Emurian (2005). After examining the website screenshot for at least four seconds, participants were asked, as in study 1, to rate the website with the *TrustDiff*, the Likert-type trust scale from Flavián et al. (2006), the UMUX- (Borsci et al., 2015) and the VisAWI-scale (Moshagen & Thielsch, 2010).

Results. ***Results of study 1: factor analysis, convergent validity, and discriminant validity.*** The item analysis and reduction process followed three steps. First, in analyzing the distribution statistics for each of the 20 items, three items were excluded due to a ceiling effect. Second, an exploratory factor analysis was conducted on the 17 remaining items. Three more items were thus eliminated because they failed to meet the minimum criteria of having a primary factor loading of .4 or above, and no cross-loading of .3 or above (Howard, 2016). A second exploratory factor analysis of the 14 remaining items showed that the three factors explained 74% of the variance and that all items had primary factor loadings of .5 or above (see Table 6). Finally, assessing reliability of each subscale, benevolence ($\alpha = .89$), integrity ($\alpha = .95$), and competence ($\alpha = .93$) showed high internal consistency. The *TrustDiff* correlates strongly ($r = .68$) with the trust questionnaire from Flavián et al. (2006), thus indicating convergent validity. Comparatively

lower correlations with visual aesthetics ($r = .46$) and usability ($r = .5$) indicate discriminant validity.

Results of study 2: confirmatory factor analysis. To test the three-dimensional factor structure, a confirmatory factor analysis was conducted. Since the goal was to create an economic scale for user trust and some items showed cross-loadings or low loadings to their factor, the scale was reduced to 10 items. The 10-item scale, measuring three related but distinct subdimensions, showed excellent psychometric properties [$\chi^2(32) = 32.500$, $p = .442$, $\chi^2/df = 1.02$, $CFI = 1.000$, $SRMR = .027$, $RMSEA = .007$, $PCLOSE = .996$] (see Figure 2) and high internal consistency ($\alpha Ben = .85$, $\alpha Int = .90$, $\alpha Com = .91$). Thus, the questionnaire could be improved and shortened without losing reliability, presenting excellent fit with high internal consistency.

Results of study 3: criterion validity. Participants viewed the website on average for 1.47 minutes ($SD = 1.4$, min = 13.8 seconds, max = 14.08 minutes). No significant differences in viewing time (log-transformed) were observed between the conditions ($t(246.88) = 0.073065$, $p = 0.9418$). Criterion validity was investigated

Table 6
Results of the second exploratory factor analysis in study 1. Three factors explain 74% of the total variance. Factor loadings above .3 are marked in bold.

	Factor loadings			
Item	Benevolence	Integrity	Competence	h2
BEN1: ignoring - caring	**.785**	.081	.059	.779
BEN2: malicious - benevolent	**.605**	.174	.058	.611
BEN4: insensitive - sensitive	**.825**	.005	-.014	.675
BEN5: inconsiderate - empathic	**.903**	-.025	.009	.790
INT1: dishonest - honest	.143	**.877**	-.113	.834
INT4: unbelievable - believable	.074	**.709**	.060	.657
INT5: untruthful - truthful	-.011	**.762**	.121	.714
INT6: fraudulent - credible	-.030	**.770**	.173	.774
COM2: incompetent - competent	-.040	.121	**.822**	.793
COM3: unskilled - skillful	.087	-.065	**.836**	.700
COM4: unqualified - proficient	-.002	.065	**.827**	.762
COM5: incapable - capable	-.051	.097	**.847**	.795
COM6: uninformed - informed	-.014	.055	**.841**	.764
COM9: inept - resourceful	.114	-.142	**.871**	.693
Eigenvalues	0.70	1.80	8.62	
% of variance	18	18	38	
α	.90	.92	.95	

Note. Total $N = 601$

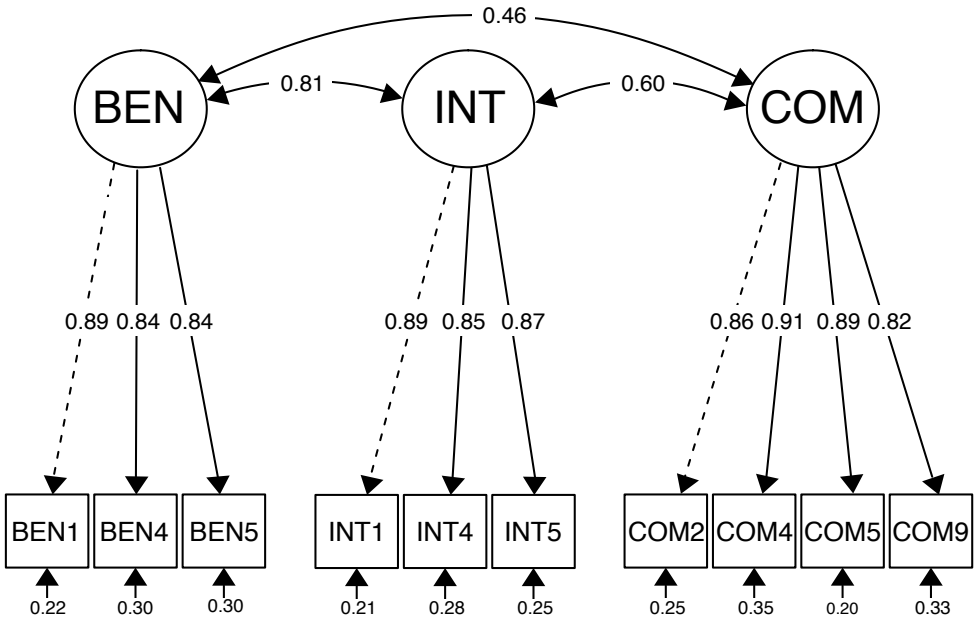


Figure 2. Measurment model of the TrustDiff in study 2 with standardized loadings. Dotted lines indicate loadings that were constrained to one [$\chi^2(32) = 32.500$, $p = .442$, $\chi^2/df = 1.02$, $CFI = 1.000$, $SRMR = .027$, $RMSEA = .007$, $PCLOSE = .996$]

using Welch's two sample t-test, comparing the high trust condition with the neutral condition. Table 7 shows that the test yielded significant differences for all subscales of the *TrustDiff* and its total score, thus indicating that the *TrustDiff* is able to differentiate between the two conditions.

Discussion and conclusion. The aim of the above depicted studies was to construct and validate a common, economic, easy-to-use and easy-to-translate scale measurement for users' trust on the web using a semantic differential, as this was a much-needed instrument that had yet to be developed (Y. Kim & Peterson, 2017). The scale was developed by following the framework of Verhagen et al. (2015), conducting three independent studies and thus generating the final item pool. In a first step, 28 positive adjectives with up to three antonyms were generated based on existing literature and the three subdimensions of trust. Following testing for appropriate linguistic and psychological bipolarity by an expert panel, the item pool was reduced to 20 item pairs. The exploratory factor analysis from study 1 further suggested that the item pool should be reduced to 14 item pairs, measuring three distinct but related subdimensions of trust. Convergent and discriminant validity

Table 7
Descriptive statistics and results of Welch's two samples t-test as an assessment of criterion validity of the TrustDiff.

		High trust (n = 128)		Neutral (n = 124)					
		M	SD	M	SD	t	df	p	d
TrustDiff	Benevolence	5.01	0.962	4.28	1.075	5.681	245.0	<.001	0.72
	Integrity	5.48	0.993	4.75	1.199	5.210	238.7	<.001	0.66
	Competence	5.58	1.011	4.47	1.455	6.989	218.6	<.001	0.89
	Total	5.37	0.899	4.50	1.176	6.577	230.1	<.001	0.84
Trust		5.11	0.947	4.20	1.254	6.470	228.8	<.001	0.82
VisAWI		4.91	1.094	3.91	1.167	7.037	247.7	<.001	0.89

Note. Total $N = 252$

were given, since the 14 items *TrustDiff* correlated relatively high with a Likert-type scale for trust (Flavián et al., 2006) but was less pronounced with visual aesthetics (Moshagen & Thielsch, 2010) and usability (Borsci et al., 2015). The confirmatory factor analysis from study 2 suggested that the item pool should be reduced to 10 items, indicating that the scale could further be shortened and improved without losing reliability while still showing excellent psychometric properties. Study 3 demonstrated the *TrustDiff*'s criterion validity by highlighting its ability to differentiate between website mockups with different numbers of trust related cues. The rating differences between the websites were between d = 0.66 and 0.89, commonly interpreted as between moderate to large (Cohen, 1988).

The main contribution of the *TrustDiff* lies in its broader applicability for measuring trust on the web. Compared to preexisting, context- and language-specific questionnaires (e.g., Flavián et al., 2006; Lu et al., 2012; McKnight et al., 2002a), the *TrustDiff* may be applied easily in different contexts. The pairs of antonyms used in the *TrustDiff* fit to various contexts related to user trust on the web. It can further be assumed that the item pairs, which solely comprise two adjectives, are easier and less time consuming to translate into other languages than the full statements of Likert-type scales (e.g., Bhattacherjee, 2002; Cho, 2006; Flavián et al., 2006; Gefen, 2002; McKnight et al., 2002a). Another advantage of semantic differentials is that the range of possible answers is more exhaustive (Chin et al., 2008). Therefore, the *TrustDiff* could be applied to investigate how different web design elements relate to different dimensions of trust and distrust. As indicated by Seckler et al. (2015), some design elements relate to distrust rather than trust.

However, despite its excellent psychometric properties it is important to apply the *TrustDiff* to various products and in different (research) contexts, although the scale demonstrated good results in that regard with various technologies described in the open answers from study 2. Since the wording of the *TrustDiff* is not exclusive to the web context, the validity could additionally be examined, for example, in offline buyer-seller relationships. Furthermore, despite the promising advantages of semantic differentials concerning their applicability in different languages, the structural validity and psychometric bipolarity of the *TrustDiff* still need to be systematically tested in other languages and cultural contexts.

Conclusion. All things considered, the advantage of the *TrustDiff* over preexisting Likert-type trust scales lies in its broader and simpler application in different contexts and languages, keeping the possible losses of reliability and validity to a minimum, and its ability to measure different manifestations of trust from a negative to a positive pole in one scale. The *TrustDiff* therefore represents a viable alternative to preexisting Likert-type questionnaires for user trust. The development and validation with over 1000 participants in three independent studies followed best practices and the scale is ready to be applied to a variety of research questions.

General Discussion

The three manuscripts that make up this dissertation represent three challenges of modern HCI-research, as depicted in the general introduction. Moreover, the studies presented in this frame extended the knowledge within their academic domain, as they uncovered new vital measuring dimensions, measuring instruments, and as they yielded new strategies and recommendations for future online research in HCI. In addition to the results, implications, and limitations, which were discussed in the corresponding sections of the three manuscripts, some final general remarks and conclusions can be made, referring to the three challenges identified in the general introduction. It is important to note, however, that some of these remarks merely represent observations made during the course of the projects for this dissertation and future research to obtain empirical evidence must be conducted in order to draw further conclusions.

Challenge 1: From Open Answers to a Quantifiable Experience - Conclusions and Future Research

The first remarks and conclusions are made concerning the first challenge, which revolved around the need to explore and uncover important dimensions in HCI-research in the light of a lack of sufficient theoretical frameworks. The aim of the study in the first manuscript was to explore the roles of high challenges and avatar death in positive player experiences. Despite the challenge-skill balance as a theoretical paradigm (Csikszentmihalyi, 1990) and the existence of several quantitative measures for this domain (IJsselsteijn et al., 2008; Oliver & Bartsch, 2010; Ryan et al., 2006; Schmierbach et al., 2014), the theoretical basis and general perception of challenge in video games were somewhat incompatible with the fact that highly challenging games with numerous occurrences of avatar death may be perceived as enjoyable positive experiences. Thus, the empirical knowledge base was not yet sufficient to fully address this situation.

A mixed-method approach, namely, the critical incident method and thematic analysis protocol (Braun & Clarke, 2006; Flanagan, 1954), was used to explore further dimensions which might come into play in a setting of excessive gameplay challenges. As described in the results and discussion section of the first manuscript, the study's main contribution was identifying moments of *achievements & victories* and *learning & improvement* as important factors for a positive player experience in

a game with excessive challenges and numerous occurrences of avatar death. The explorative mixed-method approach represented a substantial step for capturing these moments in the participants' narrations and quantifying them in order to assess their predictive power for positive player experiences. The participants' qualitative open answers turned out to be a vital part of the study for identifying such predictive factors, as none of the quantitative measures were found to be significant in this model, further highlighting that standardized scales deriving from preexisting theoretical paradigms were not fully able to address the study's aim.

Another reason for the limited value of the quantitative measures in this study possibly lies in the very specific sample used for this study: In order to explore why some players enjoy excessively challenging games, we recruited participants specifically from *Dark Souls* fan forums. It is thus unsurprising that these participants generated highly skewed distributions with sometimes little variance for certain variables such as positive affect, competence, autonomy, challenge, challenge-skill balance, and enjoyment. A prime example is the variable "enjoyment", measured with a scale developed and validated by Oliver and Bartsch (2010). We found that both groups, participants who reported positive experiences and the ones reporting negative experiences, rated the game very high on enjoyment ($Mdn = 7$, respectively 6.8). Although this difference was statistically significant, this quantitative variable was unable to reveal a meaningful and relevant result regarding the study's aim to examine the roles of high challenges and avatar death. We concluded that these results demonstrated certain limitations of Likert-type survey data in regard to our specific sample of fans, as these participants most likely tended to rate their affection towards this game as a whole and not in regard to the previously depicted open-answer experience, even though our task clearly instructed them to do so. This notion derives from Esses and Maio (2002), who claimed that participants who are asked to respond to closed-ended measures concerning specific attitude objects may at times merely report their overall attitudes. This effect may be caused by the inference of past experiences and behaviors (Bem, 1972; Olson, 1992) or by the high demands of meta-judgments, which closed-ended measures require from participants (Bassili, 1996). An open-ended measure therefore is more likely to assess specific reports associated with an attitude object, rather than just retrieving the general evaluation (Esses & Maio, 2002). In short, open answers allow researchers to gather insight not only about the participants' opinions, but also whether the participants actually understood the question and about the quality of their response.

Hence, in the case of the specific sample used for this study, an in-depth qualitative description of experience reports provided more valuable insight and a mixed-method approach proved to be a viable tool for addressing the aim of this study. Moreover, the newly identified dimensions could subsequently be quantitatively analyzed and thus provide an ideal position for future research revolving around quantitative approaches. The results of this study therefore strongly support the notion of applying open questions in online surveys for HCI-research, as they enable participants to provide answers beyond predetermined answer categories and to indicate dimensions which are not covered by closed-ended scales (Allison et al., 2002). Furthermore, open-ended questions are simple and easy to complete and they are less sample- or culture-specific, since participants answer questions in their own words (Esses & Maio, 2002).

However, there are also certain limitations attached to this mixed-method approach, especially open answers as a tool for identifying players' motivations. It should be noted that open answers also require more effort than selecting from a list of options (Holland & Christian, 2009) and they might disadvantage participants with less verbal intelligence or ability in expressing themselves in written form (Esses & Maio, 2002). Furthermore, the answers provided by participants could be affected by individual differences in conscientiousness, social desirability concerns, participants' recall ability and cognitive-emotional needs (Cacioppo, Petty, & Feng Kao, 1984; Esses & Maio, 2002; Maio & Esses, 2001). Previous research has shown that less educated respondents give a greater number of inappropriate answers to open-ended questions (Schuman & Presser, 1979, 1996), resulting in larger differences between the results of closed-ended and open-ended measures compared to participants with better education. In conclusion, the analysis of open-ended answers will ultimately favor participants with higher verbal intelligence and conscientiousness and therefore possibly lead to biased results.

The results provided in the first manuscript therefore have to be validated in further experimental studies that develop and apply standardized measures for the newly identified dimensions. Another drawback of the mixed-method approach and open-ended questions lies in the extensive, time-consuming coding and the fact that coders have to be trained for several hours before they score protocols (Allison et al., 2002; Reja et al., 2003). Closed-ended measures also allow for certain thematic distinctions in order to be more precise on various scales, while open-ended answers might have a lack of clarity in certain topics (Reja et al., 2003).

Challenge 2: Data Quality From Crowdsourced Online Samples - Conclusions and Future Research

The findings discussed in the second manuscript refer to the second challenge identified in the general introduction, namely, warranting good data quality from crowdsourced online samples. Despite the increasing popularity of online data collection owing to the advent of crowdsourcing platforms (Kan & Drummey, 2018), there is a lack of systematic analyses of data quality from crowdsourced samples. Preexisting studies analyzed other types or mixed forms of online samples (Maniaci & Rogge, 2014; Meade & Craig, 2012) or they applied an insufficient variety of measures (Hauser & Schwarz, 2016; Kan & Drummey, 2018; Peer et al., 2017) for a full assessment and adequate recommendations concerning crowdsourced data collection. The latent profile analysis in the second manuscript, taking several planned detection methods and post hoc measures into account, revealed that a group of roughly 50% of all participants answers carelessly or inattentively in a crowdsourced sample. Although it is debatable whether such a class analysis is too restrictive for excluding inattentive participants, this number is a cause for concern, as even considerably less careless responding in samples can lead to a variety of detrimental effects for a study, such as Type I (Oppenheimer et al., 2009) or Type II errors (Huang et al., 2015). Important steps and recommendations for addressing this challenge are given in the second manuscript, as we strongly encourage researchers to apply a combination of planned detection methods, post hoc measures of consistency and task-specific measures, such as assessing open answer quality. However, considering the representativeness of a sample, one could argue that a sample as presented in the study should not be used at all, as a percental exclusion of participants this high would seriously affect the generalizability of the results.

However, as revealed in other studies on careless behavior in online samples (Maniaci & Rogge, 2014; Meade & Craig, 2012), the prevalence of careless or inattentive behavior not only varied within the study of the second manuscript, depending on the detection measures used, but also differed significantly between all the studies of this dissertation. Possible reasons for this variety in data quality and the factors that may come into play will now be discussed.

In the case of the study presented in the second manuscript, the estimations for inattentive or careless participants were considerably higher than in other works (Maniaci & Rogge, 2014; Meade & Craig, 2012; Peer et al., 2017) and the reasons for

that are still of a speculative nature and subject to future research. While it seems that a crowdsourced sample is comparatively more prone to careless and inattentive behavior than student pools or mixed forms of online samples (Maniaci & Rogge, 2014; Meade & Craig, 2012), the various differences between these samples provide insights for further possible factors influencing data quality. One possible factor lies in the differing community management of crowdsourcing platforms. While *Figureeight*, which was used for data collection in our study, comprises relatively unsupervised workers with a lack of quality control mechanisms, *MTurk*, which was partially used in the study conducted by Maniaci and Rogge (2014), imposes comparatively restrictive rules and ratings on its users. Indeed, Peer et al. (2017) found the data quality on *MTurk* to be substantially better than on *Figureeight*, though this study applied attention checks only to draw this conclusion and did not use a variety of different measures.

The data cleaning process from the studies of the third manuscript, concerning the development and validation of the *TrustDiff*, does not however necessarily support the notion that data quality on *MTurk* might generally be better than on *Figureeight*. Applying the long-string index and a self-reporting item, and assessing the overall completion time, 15.8% of all participants were excluded in study 1 of the third manuscript, which used a crowdsourced sample from *MTurk*. However, the long-string index and self-report item (about whether to use the participant's data) in the second manuscript flagged 6% and 5% of all participants respectively, while overall duration time was unable to distinguish between participants with deficient and sufficient data quality. Future research should therefore aim at systematically comparing samples from different crowdsourcing platforms (e.g. *Figureeight* and *MTurk*), using various measures for detecting carelessness or inattention in order to analyze the quality differences between platforms.

The first manuscript of this dissertation provides a further notion of what might generally influence data quality in an online sample. While the highly specific sample of participants recruited from various fan forums of the game *Dark Souls* might have led to certain limitations in the interpretation of the quantitative data (see general discussion of challenge 1), the data quality was considerably better compared to the study of the second manuscript in terms of carelessness. Roughly 89% of all participants provided good or excellent open answer quality, while only 75% did so in the data quality study. Furthermore, the longstring index exclusion of 5% also provided a slightly better result compared to roughly 6.3% from the

data quality paper. Again, these notes have to be handled with caution as they do not represent a systematic comparison between these two samples. However, this might hint at a possible heightened motivation and passion for the study's subject, as the fans recruited for the first manuscript seemed to be eager to provide lengthy experience reports: The open-answer reports quantitatively surpassed the word counts from the data quality study, even though both studies instructed the participants to "provide at least 50 words".

The notion that participants' motivation and interest influences data quality has already been taken on by Meade and Craig (2012). Moreover, nonresponses in particular are usually associated with participants who were less interested in the topic of the question (Groves et al., 2004; Holland & Christian, 2009): A study conducted by Holland and Christian (2009) concluded that participants who were very interested in a topic have significantly fewer nonresponses and their responses are of higher quality (more words, topics and elaboration) than participants who were merely somewhat interested. Another study found that people's interest in the survey topic also has a greater impact on response rates than other features, including incentives provided to respondents (Cook, Heath, & Thompson, 2000). Whether and how strongly motivational factors or interest in a certain topic do indeed improve data quality in crowdsourced samples is, however, also subject to systematic, future research analyzing the correlation between participants' interest and various measures of data quality. As suggested in the general introduction, data quality in crowdsourced samples and online samples in general depends on a multitude of factors, which should be taken into consideration when planning to recruit participants from such platforms and, aside from the detection of carelessness or inattention in online samples, the prevention of poor data quality is also still less well understood (Clifford & Jerit, 2015; Meade & Craig, 2012). Researchers should take all these factors into consideration when collecting their survey data online, as detecting carelessness alone might not fully solve the problem if one has to exclude half of all participants in the end. The results discussed in the second manuscript, as well as the data quality assessments from all papers in this dissertation, indicate that good or poor data quality stems from a multitude of factors and that online data collection requires a lot of experience, knowledge and careful planning.

Challenge 3: Applying Common Standardized and Validated Measures – Conclusions and Future Research

The third challenge introduced in this dissertation revolves around the lack of a common theoretical definition and the subsequent lack of a validated and easy-to-use measure for trust on the web – a situation which emerged from numerous variations and discord in regard to the dimensional concept of trust (Van der Werff et al., 2018) and the variety of specific contexts where trust was measured using tailor-made instruments (Cho, 2006; Flavián et al., 2006). The studies discussed in the third manuscript tackled this challenge by developing and validating a semantic differential for measuring trust on the web. The *TrustDiff* demonstrated excellent psychometric scales and promises a viable alternative for other Likert-type trust scales. Moreover, the differential is easily applicable in various contexts and easier to translate into different languages. As mentioned in the relevant discussion section however, the promising characteristics of this scale still have to be validated in future studies conducted in different contexts.

The lack of a common validated measure for trust on the web is exemplary, as the introduction of the UX notion in HCI-research is still relatively new (Law et al., 2009) and as technical advancements are fast moving, possibly changing many concepts within a short period of time, which is highlighted by the discussion about trust in technological trustees (Van der Werff et al., 2018). In that manner, the third challenge discussed is somewhat reminiscent of the first, as the domain surrounding challenges in video games also lacked a clear theoretical basis and relevant measures to apply. However, while the basic paradigms from challenges in video games were insufficient to provide any possible explanation for the phenomena identified in the first manuscript, the preexisting literature surrounding trust on the web was not characterized by the lack of explanatory paradigms but rather by an overabundance of definitions (Van der Werff et al., 2018). Thus, in the first challenge the goal was primarily to explore further important dimensions of this domain, whereas the third challenge involved a standardization of theoretical concepts, relevant measurements, and items.

This showcases that the three challenges for HCI-research discussed in this dissertation are, however, intertwined and a clear-cut recommendation may therefore be difficult to formulate. Depending on the exact aims of a study, it would for example also be legitimate to develop and validate a new measure from a situ-

ation with a certain lack of theoretical background. Contrariwise, the inclusion of open-ended measures may also be a valuable tool for domains with an overabundance of theoretical definitions. As indicated in the second manuscript, however, the implementation of open questions generates a whole new subgroup of careless participants and assessing this task-specific answer quality is a necessity in filtering out careless respondents. Hence, the second challenge surrounding the issue of warranting good data quality is a ubiquitous issue in the other two challenges as well. Since both manuscripts, the one addressing the first challenge and the one concerning the third challenge, included open questions in order to achieve their goals, the hurdles and difficulties of warranting good data quality, as discussed in the second manuscript, also apply to these studies. In particular, if the exploration of new dimensions fully relies on open answers, circumstances which foster good answer quality should be taken into consideration, as low motivation and disinterested participants might jeopardize such an enterprise. In terms of the development of a new scale, as depicted in the third manuscript, previous studies have shown that insufficient data quality may also cause problems in the item analysis and in investigating questionnaire dimensionality (Johnson, 2005), further indicating the interdependence of these challenges. Therefore, the three challenges should not be viewed in isolation but rather as basic themes which may simultaneously occur in a domain of research that is characterized by rapid technological advancements and the continuous introduction of new concepts.

Conclusion

The manuscripts of this dissertation highlighted three challenges that are part of a relatively new and fast-moving domain of research, sometimes characterized by insufficient theoretical groundwork or a lack of common and validated concepts, including relevant measures. Research in HCI is further complicated by the increasing trend of online data collection which is accompanied by a possible corruption of the data quality if online surveys are not carefully planned and the data quality assessed based on empirical findings and recommendations.

The first manuscript depicts the way in which an explorative, mixed-method approach revealed new dimensions of interest in a domain with previously insufficient theoretical frameworks. The qualitative data turned out to be a vital part of the study for predicting positive player experiences with highly challenging gameplay, as none of the preestablished quantitative measures had any predictive power

in this model. The study hence showcased the limitations of standardized scales and demonstrated the advantages that go along with open-ended questions. The second manuscript comprises a guide and recommendations for the assessment of data quality from crowdsourced online surveys. The results suggest that this type of online collected data needs to be assessed carefully, taking several planned detection methods, post hoc measures, and task specific measures into consideration. Moreover, data quality of online samples seems to be highly dependent on a multitude of contextual and motivational factors, which additionally need to be taken into account. Lastly, the third manuscript describes the development and validation of a measure for trust on the web, a domain which previously lacked common and validated concepts and measures. As technical advancements in the field of HCI are fast moving, versatile and easy-to-translate measures are of vital importance for conducting research. The third study was able to address this challenge by establishing a semantic differential for measuring trust on the web.

These challenges indicate that exploring new fields of research requires not only theoretical knowledge of a domain but also experience in applying a variety of methodological approaches. The manuscripts presented in this dissertation aim at providing new insights, strategies, and recommendations for other researchers for managing and overcoming these challenges of modern HCI-research.

References

Alexander, J. T., Sear, J., & Oikonomou, A. (2013). An investigation of the effects of game difficulty on player enjoyment. *Entertainment Computing*, *4*(1), 53–62.

Allison, F., Carter, M., & Gibbs, M. (2015). Good frustrations: The paradoxical pleasure of fearing death in DayZ. In *Proceedings of the annual meeting of the australian special interest group for computer human interaction* (pp. 119–123).

Allison, L. D., Okun, M. A., & Dutridge, K. S. (2002). Assessing volunteer motives: A comparison of an open-ended probe and likert rating scales. *Journal of Community & Applied Social Psychology*, *12*(4), 243–255.

Aponte, M.-V., Levieux, G., & Natkin, S. (2011). Difficulty in videogames: An experimental validation of a formal definition. In *Proceedings of the 8th international conference on advances in computer entertainment technology* (p. 49).

Bart, Y., Shankar, V., Sultan, F., & Urban, G. L. (2005). Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study. *Journal of Marketing*, *69*(4), 133–152.

Bartle, R. A. (2004). *Designing virtual worlds*. San Francisco, CA: New Riders.

Bassili, J. N. (1996). Meta-judgmental versus operative indexes of psychological attributes: The case of measures of attitude strength. *Journal of Personality and Social Psychology*, *71*(4), 637–653.

Bateman, S., Mandryk, R. L., Stach, T., & Gutwin, C. (2011). Target assistance for subtly balancing competitive play. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2355–2364).

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York, NY: Academic Press.

Bente, G., Rüggenberg, S., Krämer, N. C., & Eschenburg, F. (2008). Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human Communication Research*, *34*(2), 287–318.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems*, *19*(1), 211–241.

Bopp, J. A., Mekler, E. D., & Opwis, K. (2016). Negative emotion, positive experience?: Emotionally moving moments in digital games. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2996–3006).

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction*, *31*(8), 484–495.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307.

Carter, M., Gibbs, M., & Wadley, G. (2013). Death and dying in DayZ. In *Proceedings of the 9th australasian conference on interactive entertainment: Matters of life and death* (pp. 1–6).

Casalo, L. V., Flavián, C., & Guinalíu, M. (2007). The role of security, privacy, usability and reputation in the development of online banking. *Online Information Review*, *31*(5), 583–603.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160.

Cechanowicz, J. E., Gutwin, C., Bateman, S., Mandryk, R., & Stavness, I. (2014). Improving player balancing in racing games. In *Proceedings of the first ACM SIGCHI annual symposium on computer-human interaction in play* (pp. 47–56).

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, *26*(7), 1131–1139.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*(1), 53–81.

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31–34.

Chen, S. C., & Dhillon, G. S. (2003). Interpreting dimensions of consumer trust in e-commerce. *Information Technology and Management*, *4*(2-3), 303–318.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361.

Chin, W. W., Johnson, N., & Schwarz, A. (2008). A fast form approach to measuring technology acceptance and other constructs. *MIS Quarterly*, *32*(4), 687–703.

Cho, J. (2006). The mechanism of trust and distrust formation and their relational outcomes. *Journal of Retailing*, *82*(1), 25–35.

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, *79*(3), 790–802.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum Associates.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836.

Copcic, A., McKenzie, S., & Hobbs, M. (2013). Permadeath: A review of literature. In

Games innovation conference (IGIC), 2013 IEEE international (pp. 40–47).

Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: A study of consumer perceptions. *Electronic Commerce Research and Applications*, *2*(3), 203–215.

Csikszentmihalyi, M. (1990). *Flow. The psychology of optimal experience.* New York, NY: Harper Perennial.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

Deutsch, M. (1962). *Cooperation and trust: Some theoretical notes.* Lincoln, NE: University of Nebraska Press.

De Winter, J., Kyriakidis, M., Dodou, D., & Happee, R. (2015). Using crowdflower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing*, *3*, 2518–2525.

Diekmann, A. (2009). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (18th ed.). Reinbek, DE: Rowohlt.

Dietz, G., & Den Hartog, D. N. (2006). Measuring trust inside organisations. *Personnel Review*, *35*(5), 557–588.

Dogan, V. (2018). A novel method for detecting careless respondents in survey data: Floodlight detection of careless respondents. *Journal of Marketing Analytics*, *6*(3), 95–104.

Erikson, E. H. (1963). *Childhood and society* (2nd ed.). New York, NY: Norton.

Esses, V. M., & Maio, G. R. (2002). Expanding the assessment of attitude components and structure: The benefits of open-ended measures. *European Review of Social Psychology*, *12*(1), 71–101.

Falstein, N. (2005). Understanding fun – the theory of natural funativity. In S. Rabin (Ed.), *Introduction to game development* (pp. 71–98). Boston, MA: Charles River Media Hingham.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4), 327–358.

Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, *43*(1), 1–14.

Flynn-Jones, E. (2015). Don't forget to die: A software update is available for the death drive. In T. E. Mortensen, J. Linderoth, & A. M. Brown (Eds.), *The dark side of game play* (pp. 58–72). Abingdon-on-Thames, UK: Routledge.

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, *40*(5), 873–884.

Friedman, B., Khan Jr, P. H., & Howe, D. C. (2000). Trust online. *Communications of the ACM*, *43*(12), 34–40.

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 1631–1640).

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-learning and Digital Media*, *2*(1), 5–16.

Gefen, D. (2002). Reflections on the dimensions of trust and trustworthiness among online consumers. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *33*(3), 38–53.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*, 877–902.

Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *68*(1), 2–31.

Gutwin, C., Rooke, C., Cockburn, A., Mandryk, R. L., & Lafreniere, B. (2016). Peak-end effects on player experience in casual games. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5608–5619).

Hassanein, K., & Head, M. (2007). Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping. *International Journal of Human-Computer Studies*, *65*(8), 689–708.

Hassenzahl, M. (2008). User experience (UX): Towards an experiential perspective on product quality. In *Proceedings of the 20th conference on l'interaction homme-machine* (pp. 11–15).

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In J. Ziegler & G. Szwillus (Eds.), *Mensch & Computer 2003. Interaktion in Bewegung* (pp. 187–196). Stuttgart, Leipzig, DE: B.G. Teubner.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.

Hawkins, D. I., Albaum, G., & Best, R. (1974). Stapel scale or semantic differential in marketing research? *Journal of Marketing Research*, *11*(3), 318–322.

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review*, *27*(2), 196–212.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845.

Hunicke, R. (2005). The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI international conference on advances in computer entertainment technology* (pp. 429–433).

IJsselsteijn, W., Van Den Hoogen, W., Klimmt, C., De Kort, Y., Lindley, C., Mathiak, K., ... Vorderer, P. (2008). Measuring the experience of digital game enjoyment. In *Proceedings of measuring behavior* (pp. 88–89).

International Organization for Standardization. (2010). *ISO 9241-210:2010 Ergonomics of human-system interaction: Part 210: Human-centred design for interactive systems.* Geneva, CH: International Organization for Standardization (ISO).

Jarvenpaa, S. L., Tractinsky, N., & Saarinen, L. (1999). Consumer trust in an internet store: A cross-cultural validation. *Journal of Computer-Mediated Communication*, *5*(2), 1–35.

Jin, S.-A. A. (2012). "Toward integrative models of flow": Effects of performance, skill, challenge, playfulness, and presence on flow in video games. *Journal of Broadcasting & Electronic Media*, *56*(2), 169–186.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory — versions 4a and 54.* Berkeley, CA: University of California, Institute of Personality and Social Research.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129.

Juul, J. (1999). *A clash between game and narrative: A thesis on computer games and interactive fiction.* Copenhagen, DK: University of Copenhagen.

Juul, J. (2009). Fear of failing? The many meanings of difficulty in video games. In B. Perron & M. J. Wolf (Eds.), *The video game theory reader 2* (p. 237-252). New York, NY: Routledge.

Juul, J. (2011). *Half-real: Video games between real rules and fictional worlds.* Cambridge, MA: MIT press.

Juul, J. (2013). *The art of failure: An essay on the pain of playing video games.* Cambridge, MA: MIT Press.

Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Computers in Human Behavior*, *83*, 243–253.

Kim, J., & Moon, J. Y. (1998). Designing towards emotional usability in customer interfaces — trustworthiness of cyber-banking system interfaces. *Interacting with Computers*, *10*(1), 1–29.

Kim, Y., & Peterson, R. A. (2017). A meta-analysis of online trust relationships in e-commerce. *Journal of Interactive Marketing*, *38*, 44–54.

Klarkowski, M., Johnson, D., Wyeth, P., McEwan, M., Phillips, C., & Smith, S. (2016). Operationalising and evaluating sub-optimal and optimal play experiences through challenge-skill manipulation. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5583–5594).

Klastrup, L. (2006). Death matters: Understanding gameworld experiences. In *Proceedings of the 2006 ACM SIGCHI international conference on advances in computer entertainment technology.*

Klimmt, C., Blake, C., Hefner, D., Vorderer, P., & Roth, C. (2009). Player performance, satisfaction, and video game enjoyment. In *International conference on entertainment computing* (pp. 1–12).

Koufaris, M., & Hampton-Sosa, W. (2004). The development of initial trust in an online company by new customers. *Information & Management*, *41*(3), 377–397.

Lankton, N. K., & McKnight, D. H. (2011). What does it mean to trust facebook?: Examining technology and interpersonal trust beliefs. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *42*(2), 32–54.

Lauer, T. W., & Deng, X. (2007). Building online trust through privacy practices. *International Journal of Information Security*, *6*(5), 323–331.

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. In *Mensch & Computer* (pp. 125–134).

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 719–728).

Lazzaro, N. (2004). Why we play games: Four keys to more emotion without story. In *Presentation at the game developers conference San Jose* (pp. 1–8).

Lee, M. K., & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce*, *6*(1), 75–91.

Lewicki, R. J., & Brinsfield, C. (2012). Measuring trust beliefs and behaviours. In F. Lyon, G. Moellering, & M. N. Saunders (Eds.), *Handbook of research methods on trust* (pp. 29–39). Cheltenham, UK: Edward Elgar.

Lu, J., Wang, L., & Hayes, L. A. (2012). How do technology readiness, platform functionality and trust influence C2C user satisfaction? *Journal of Electronic Commerce Research*, *13*(1), 50–69.

Maio, G. R., & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of Personality*, *69*(4), 583–614.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

McEvily, B., Perrone, V., & Zaheer, A. (2003). Trust as an organizing principle. *Organization*

Science, *14*(1), 91–103.

McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior*, *84*, 295–303.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 1–15.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002a). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002b). The impact of initial consumer trust on intentions to transact with a web site: A trust building model. *The Journal of Strategic Information systems*, *11*(3-4), 297–323.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review*, *23*(3), 473–490.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Mirnig, A. G., Meschtscherjakov, A., Wurhofer, D., Meneweger, T., & Tscheligi, M. (2015). A formal analysis of the ISO 9241-210 definition of user experience. In *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems* (pp. 437–450).

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709.

Naquin, C. E., Kurtzberg, T. R., & Belkin, L. Y. (2010). The finer points of lying online: E-mail versus pen and paper. *Journal of Applied Psychology*, *95*(2), 387–394.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11.

Oliver, M. B., & Bartsch, A. (2010). Appreciation as audience response: Exploring entertainment gratifications beyond hedonism. *Human Communication Research*, *36*(1), 53–81.

Olson, J. M. (1992). Self-perception of humor: Evidence for discounting and augmentation effects. *Journal of Personality and Social Psychology*, *62*(3), 369–377.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Pavlou, P. A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research*, *15*(1), 37–59.

Pavlou, P. A., Liang, H., & Xue, Y. (2007). Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective. *MIS Quarterly*, *31*(1), 105–136.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Peng, W., Lin, J.-H., Pfeiffer, K. A., & Winn, B. (2012). Need satisfaction supportive game features as motivational determinants: An experimental study of a self-determination theory guided exergame. *Media Psychology*, *15*(2), 175–196.

Prendinger, H., Puntumapon, K., & Madruga, M. (2016). Extending real-time challenge balancing to multiplayer games: A study on eco-driving. *IEEE Transactions on Computational Intelligence and AI in Games*, *8*(1), 27–32.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics*, *19*(1), 160–117.

Ritterfeld, U., Cody, M., & Vorderer, P. (2009). *Serious games: Mechanisms and effects.* Routledge, Abington-on-Thames, UK.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344–360.

Schmierbach, M., Chung, M.-Y., Wu, M., & Kim, K. (2014). No one likes to lose. *Journal of Media Psychology*, *26*(3), 105–110.

Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review*, *44*(5), 692–712.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context.* Thousand Oaks, CA: Sage.

Seckler, M., Heinz, S., Forde, S., Tuch, A. N., & Opwis, K. (2015). Trust and distrust on the web: User experiences and website characteristics. *Computers in Human Behavior*, *45*, 39–50.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using mechanical turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220.

Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, *80*(2), 325–339.

Sherry, J. L. (2004). Flow and media enjoyment. *Communication Theory*, *14*(4), 328–347.

Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology*, *57*, 529–555.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*(2), 325–337.

Söllner, M., Pavlou, P., & Leimeister, J. (2013). *Understanding trust in it artifacts – a new conceptual approach*. Available at Social Science Research Network (SSRN 2475382).

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748.

Sweetser, P., & Wyeth, P. (2005). Gameflow: A model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, *3*(3), 1–24.

Tan, C. H., Tan, K. C., & Tay, A. (2011). Dynamic game difficulty scaling using adaptive behavioural based AI. *IEEE Transactions on Computational Intelligence and AI in Games*, *3*(4), 289–301.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Tuch, A. N., Schaik, P. V., & Hornbæk, K. (2016). Leisure and work, good and bad: The role of activity domain and valence in modeling user experience. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *23*(6), 35.

Van Auken, S., & Barry, T. E. (1995). An assessment of the trait validity of cognitive age measures. *Journal of Consumer Psychology*, *4*(2), 107–132.

Van der Werff, L., Real, C., & Lynn, T. (2018). Individual trust and the internet. In R. Searle, A. Nienaber, & S. Sitkin (Eds.), *The routledge companion to trust* (pp. 1–33). Oxford, UK: Routledge.

Verhagen, T., Van Den Hooff, B., & Meents, S. (2015). Toward a better use of the semantic differential in is research: An integrative framework of suggested action. *Journal of the Association for Information Systems*, *16*(2), 108–143.

Vicencio-Moreira, R., Mandryk, R. L., & Gutwin, C. (2015). Now you can compete with anyone: Balancing players of different skill levels in a first-person shooter game. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 2255–2264).

Vorderer, P., Klimmt, C., & Ritterfeld, U. (2004). Enjoyment: At the heart of media entertainment. *Communication Theory*, *14*(4), 388–408.

Wang, Y. D., & Emurian, H. H. (2005). An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior*, *21*(1), 105–125.

Ward, M., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417–430.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief

measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Wilson, D., & Sicart, M. (2010). Now it's personal: On abusive game design. In *Proceedings of the international academic conference on the future of game design and technology* (pp. 40–47).

Wirtz, J., & Lee, M. C. (2003). An examination of the quality and context-specific applicability of commonly used customer satisfaction measures. *Journal of Service Research*, *5*(4), 345–355.

Yun, C., Trevino, P., Holtkamp, W., & Deng, Z. (2010). PADS: Enhancing gaming experience using profile-based adaptive difficulty system. In *Proceedings of the 5th ACM SIGGRAPH symposium on video games* (pp. 31–36).

Acknowledgements

I would like to thank Prof. Dr. Klaus Opwis for his trust and support of my dissertation thesis and the related study projects. I would further like thank the PhD Committee, namely Prof. Dr. Jana Nikitin and Prof. Dr. Rainer Greifeneder, for acting as second supervisor and chair of my disputation, respectively. My further thanks go to all my co-authors, namely Florian Brühlmann, Glena Iten, Lena Aeschbach, Elisa Mekler, Denise Rieser, and Klaus Opwis, who always gave me helpful insights and constructive criticism and feedback. Lastly, I thank the entire MMI Department at the University of Basel and my family for their general support during the doctorate. This dissertation was proofread by Alexa Barnby.

Statement of Authorship

i I, Serge Petralito, hereby declare that I have written the submitted doctoral thesis "Current Challenges in HCI-Research: Quantifying Open Experiences, Warranting Data Quality, and Developing Standardized Measures" without any assistance from third parties not indicated.

ii I only used the resources indicated.

iii I marked all citations.

iv My cumulative dissertation is based on three manuscripts. The first has already been published. The second and the third manuscript are submitted. I certify here that the articles in this dissertation concern original work. I contributed substantially to all manuscripts in this dissertation and I have been jointly responsible for the idea, conception, methodological design, data collection, analyses, and writing of all manuscripts. This characterization of my contributions is in agreement with my co-authors' views.

Place and Date: _____

Serge Petralito: _____

Appendix

1. Petralito, S., Brühlmann, F., Iten, G., Mekler, E. D., Opwis, K. (2017). A Good Reason to Die: How Avatar Death and High Challenges Enable Positive Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 5087-5097). ACM.

2. Brühlmann, F., Petralito, S., Aeschbach, L. F., Opwis, K. (submitted). Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples. [Manuscript submitted to *Plos One*]

3. Brühlmann, F., Petralito, S., Rieser, D. C., Aeschbach, L. F., Opwis, K. (submitted). TrustDiff: Development and Validation of a Semantic Differentialfor User Trust on the Web. [Manuscript submitted to *International Journal of Human-Computer Studies*]

4. Curriculum Vitae, Serge Petralito

A Good Reason to Die: How Avatar Death and High Challenges Enable Positive Experiences

Serge Petralito[1], Florian Brühlmann[1], Glena Iten[1], Elisa D. Mekler[2] and Klaus Opwis[1]

[1]Center for Cognitive Psychology and Methodology, Department of Psychology, University of Basel
[2]HCI Games Group, Games Institute, University of Waterloo
{s.petralito, florian.bruehlmann, glena.iten, klaus.opwis}@unibas.ch, emekler@uwaterloo.ca

ABSTRACT

Appropriate challenges and challenge-skill balance are usually key to positive player experiences. However, some games such as the successful series *Dark Souls* are notorious for their excessive difficulty. Yet, there has been little empirical investigation of why players enjoy games they constantly struggle and fail with. We surveyed 95 participants right after the release of *Dark Souls III* about their experiences with the game, employing both open questions and different player experience measures. Players generally enjoyed challenging play sessions and mostly reported positive experiences, with achievement and learning moments strongly contributing to positive experiences. However, these factors themselves were enabled by negative events such as difficulties and avatar death. Our findings showcase that negative events bear a potential for forming positive and meaningful experiences, thus expanding previous knowledge about the role of challenge and failing in games. Moreover, the significance of hard-earned achievements extends present design conventions.

ACM Classification Keywords

J.4 Social and Behavioral Sciences: Sociology, Psychology; K.8.0 Personal Computing: Games

Author Keywords

Games; Player Experience; Failure; Challenge; Avatar Death; Enjoyment

INTRODUCTION

Video games are played for the interactive experience they provide. Challenge is ubiquitously seen as one of the most important components of this experience: an unchallenging game will likely be perceived as shallow or boring and thus might not be a particularly enjoyable experience [19, 23, 36, 38]. Although challenge and competition have been found to increase enjoyment in general [27, 34, 42], they can also be excessive and lead to negative experiences when players fail and feel less competent [35]. According to the theory of flow [11], a key characteristic of an optimal experience is the balance between challenge and skill. Hence, if challenge demands imposed by the game are too high or too low in regard to the player's skill level, playing the game leads to anxiety or boredom. The significance of an ideal challenge-skill balance is strongly emphasized in current research [3, 4, 13, 23, 33, 37], where adjustable and adaptive difficulty mechanics play an integral part in keeping this balance [8, 14, 35, 39]. Moreover, balance and accessibility represent two key notions of the *casual revolution*, a design trend towards making games more accessible by removing perceived barriers, penalties and frustrations and targeting much broader audiences than games used to over roughly a decade ago [20, 22]. In conclusion, challenge in current literature and modern game design has to a large extent been treated as a *Goldilocks factor*: The difficulty of a game should be neither too demanding nor too low in order to avoid negative experiences and frustrations.

In light of present design conventions, some exceptional games stand out, ignoring most of the conventional balancing efforts by implementing very high challenges and high consequential avatar death in their core design: For example, the popular action role-play series *Dark Souls*, which became notorious for its high difficulty as well as for its harsh penalties resulting from failure. The series in some respects acts the complete opposite way in comparison to the above mentioned trends: Challenges and demands imposed by the game are steep from the very beginning and players will constantly fail throughout the progression of the game. There are neither adaptive mechanics nor adjustable difficulty levels and still, the game series gained increasing popularity over the last years [30].

Although the appeal of punishing or unfair games has received some attention in current research [2, 27, 43], empirical evidence is yet scarce. Given the crucial role of challenge-skill balance in shaping positive experiences, we aim to examine this seemingly contradictory situation of players enjoying a game defined by high challenges, numerous frustrations and punishing avatar death. We therefore asked 95 participants to report an outstanding experience with *Dark Souls III* right after the release of the game. Participants were split in two groups, depending on whether they reported a positive or negative experience. By analyzing experience reports and employing various psychometric scales, we explored common gameplay themes and events, how they are associated with each other and how they contribute to positive experiences. Our findings suggest that negative events such as avatar death and difficulties bear a potential for enabling positive experiences. Moreover, we identified moments of learning and achievement, evoked and characterized by failures and difficulties, as being crucial

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2017, May 6-11, 2017, Denver, CO, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-4655-9/17/05 ...$15.00.
http://dx.doi.org/10.1145/3025453.3026047

for positive experiences in this context. We thus extend our notion of a positive experience by reflecting on how closely related they are to negative events.

RELATED WORK

Challenge in Dark Souls III

In *Dark Souls III* the player explores a fantasy world, consisting of continuous interconnected areas, while facing different types of enemies. The better part of its difficulty stems from engaging in combat with enemies, adapting to their varying combat patterns and finding the right strategy to defeat them. Especially larger opponents (boss fights) usually take multiple attempts in order to find out how to dodge their attacks, to identify their weak spots and defeat them. Avatar death in *Dark Souls III* is highly consequential to the player, unlike a vast majority of modern games where death merely is a minor setback [2, 10]. With failure the player loses all their souls (experience points) and respawns often at a distant place from his or her previous death, thus being forced to play a large section of the game all over again. If the player dies again before retrieving their souls from the location of their last death, all experience points are gone for good. Hence, death can result in a loss of hours of game play and immense frustration. Frequent avatar death is a core element of *Dark Souls III* due to its high difficulty.

Avatar Death

Avatar death typically serves as a game mechanic, which is used to mark the player's failure and temporary removal from play and occurs in most genres of video games [7]. The mechanic imposes a penalty on the player, consisting of repetition and incremental progress towards mastery of a certain section of the game [12]. Avatar death thus may result in an immediate negative experience, depending on how punishing the consequences for the player are. The role of a highly consequential death mechanic has been examined in two studies on the game *DayZ*, where every avatar death forces the player to restart the game from the beginning. Carter et al. [7] examined how this consequentiality affects the players' experience and behavior in the game and concluded that it leads to intensified social interactions, moral dilemmas and a raised level of perceived investment. Allison et al. [2] found similar results, concluding that the awareness of risks imbues actions with meaning and that this meaningfulness stems from a pattern of smaller negative experiences. Achievements therefore led to extremely positive emotions, because the player is aware of the high stakes the game imposes on them. Players considered the high consequentiality as a frustrating, although crucial component for the enjoyment of the game. Allison et al. [2] further stated that negative and positive affect are not mutually exclusive: the positive experience is directly created by negative feelings of fear, anxiety and unpredictability. Accordingly, avatar death is accompanied by a short-term negative affect but ultimately leads to a positive experience, if the player is able to achieve his or her goal. Avatar death and failure are often depicted as part of a learning process. Flynn-Jones [12] describes in-game death and its current loss of control as a recurring process of agency (getting back control after death), repetition and ultimately mastery of a certain section of the game. This is much in the same vein as Juul [18], stating that in-game death is the death you survive and learn from. Furthermore, death motivates players to meet challenges by learning how to overcome failure [22], thus making death and failure a recurring, informing learning process, critical to the enjoyment of a videogame [12, 18, 24, 26]. As stated by Iacovides et al. [15], challenges provide players with potential opportunities to learn, although moments of learning only occur if a breakdown (caused by challenges) leads to a breakthrough in understanding.

High challenges and enjoyment

Avatar death may play a substantial part in forming positive experiences through achievements and learning moments. However, excessive challenges are more often portrayed in literature as a negative factor for enjoyment [25, 35], which has to be throttled and adjusted to the player through adaptive difficulty mechanics [8, 14, 31, 39, 44]. Dynamic difficulty adjustments (DDA) such as auto-aim function in combat games, adaptive AI, or the rubber band adjustment in racing games help less skilled players to stay competitive and to succeed in achieving their goals [8, 14, 35, 39]. Studies show that adaptive mechanics are especially preferred by novices to feel more competent, autonomous and related to other players [8]. In addition, their implementation is perceived as a measure to make the game more enjoyable [41] and support flow [9].

A closer look at current research revolving around difficulty and enjoyment illustrates the emphasis on the importance of balanced or generally lowered difficulty settings: Schmierbach et al. [35] examined the relationship between difficulty and enjoyment and possible mediating roles of competence (as defined in self-determination-theory, [34]) and challenge-skill balance [11]. Results showed that individuals playing a harder version of a game felt less competent, reducing their sense of challenge-skill balance, which in turn diminished their enjoyment. In a similar vein, Klimmt et al. [25] found that even experienced players reported greater enjoyment and satisfaction when playing a shooter game in an easy mode, compared to medium or hard settings. Though Alexander et al. [1] put this finding into perspective by showcasing that only casual players enjoy lower difficulty settings more, regardless of their abilities, while experienced players prefer challenges according to their skill level. Furthermore, Jin [17] demonstrated that players who avoided avatar death in their play session, reported higher levels of competence and flow, although the study did not directly include measurements for enjoyment. Another study conducted by Peng et al. [29] showed that players reported greater enjoyment if the game featured adaptive difficulty mechanics, which primarily helped to reduce challenge. Klarkowski et al. [23] also emphasized the importance of an ideally balanced game, which is associated with heightened positive affect and higher enjoyment as well as heightened feelings of autonomy and relatedness, when compared to high and low challenge settings. Players furthermore seem to prefer lower levels of difficulty at the beginning of a game [25].

In conclusion, excessive challenges in games are typically associated with lower enjoyment or fun [13, 35], lowered positive and heightened negative affect as well as lowered feelings

of competence, autonomy and relatedness [23]. However, some articles also discussed the appeal of highly challenging or unfair and punishing games. Lazzaro [27] states that effort and frustrations are needed in order to reach the *fiero* state: A personal triumph over adversity by overcoming difficult obstacles. She describes it as hard fun, a type of gameplay, which revolves around achieving goals. Similar to Allison et al.'s study [2] on *DayZ*, achievements represent a central aspect in this theory. Wilson & Sicart [43] argue that unfair games can be funny *because* they are user-unfriendly and break every good-practice level-design rule. The player is in on the joke and realizes that the game designer is basically mocking him or her, while accepting this challenge as a contest between him and the creators. Games like these may thus be not functional per se but they enhance interpersonal aspects and allow for new experiences, which enable unusual ways of enjoyment.

In summary, while current research provides some valuable insights into the role of avatar death and high challenges in video games, some questions still remain open. Avatar death might play an important role for the fun in high challenging games, though these findings mostly come from theoretical works and empiric evidence is yet scarce. Moreover, literature mostly points in a direction where excessive challenges are detrimental to the players' enjoyment, even though the popularity of games like *Dark Souls* or *Super Meat Boy* contradicts or at least relativizes such findings. Therefore, the present study aims for a better understanding of the role of avatar death and high challenges in positive player experiences. The central questions are, if players enjoy high-challenging games despite occurring difficulties and failures, or if avatar death and high challenges actually may form and enable positive experiences. We thus conducted an online-survey, applying a combination of qualitative and quantitative methods. Several psychometric scales and open-ended questions were utilized in order to analyze participants' latest outstanding experiences with the game *Dark Souls III*.

METHOD

Participants of this study were asked to report an outstanding positive or negative experience in their recent play-session of *Dark Souls III*. This critical incident method allows collecting and combining qualitative and quantitative data and is focused on what the players themselves consider a positive or negative experience.

Participants

The study was distributed on several gaming forums (e.g., Steam), social networks (e.g., Facebook, Twitter, VKontakte), as well as gaming-related groups on Reddit and Facebook. The age of the 95 participants (4.2% female) who completed the survey in entirety ranged from 13 to 42 years (M = 23.81). On average, participants had been playing *Dark Souls III* for 41.83 hours (ranging from less than 1 hour to up to 346 hours) with half of the participants having played 23 hours or less. Apart from 4 participants, all respondents indicated that they had at least a little experience with any of the previous games of the series (*Dark Souls I* or *II*). In exchange for completing the survey, participants could enter a lottery to win a $100 or one of four $25 (USD) Amazon gift cards.

Procedure

The online survey consisted of both qualitative, open-ended questions, as well as several scales. The open-ended questions followed a similar approach as Bopp et al. [5]. The first open-ended question was as follows: *"Bring to mind an outstanding positive or negative experience you had in your most recent game-session in Dark Souls III"*. Additionally, they were asked to try to describe this particular experience as accurately, detailed and concrete as possible. Participants had to write down at least 50 words. Afterwards, participants were asked to clarify the cause of the thoughts and feelings they had in their experience.

After answering the open-ended questions, participants had to rate their experience in terms of positive and negative affect, need fulfillment, challenge, challenge-skill balance and enjoyment. Finally, they were asked to provide some information on demographics and experience with the previous games of the series. At the end of the survey, participants could enter their email address if they wished to participate in the prize raffle. The survey took on average 21.9 minutes to complete.

Thematic analysis of experience reports

The open-ended answers were manually coded following the thematic analysis protocol [6] to identify recurring themes. The most common themes identified were *achievements & victories, learning & improvement, difficulties & failures, lack of progress* and *enemy encounters*. The meaning of these categories is described in-depth in the results section.

The first author coded all open-ended answers. To assure interrater reliability, an independent rater coded a random subset of 41 experiences. After the category *lack of progress* was dropped, due to low agreement among the raters, a substantial agreement among the raters for all themes ($\kappa = .6$) was achieved.

Overall valence of the experience

The overall valence of the experience was coded either positive or negative on a whole-experience level. The interrater agreement was substantial ($\kappa = .65$). There were several experiences that included negative events, such as avatar death, but ultimately depicted a positive outcome, such as overcoming these challenges and reporting satisfaction. The overall valence was coded positive, if the play session was mainly described with positive outcomes such as joy, satisfaction, happiness, and positivity.

'A really outstanding experience I had during Dark Souls III would be, when I finally beat the Soul of Cinder after many attempts and the credits rolled. I can say this was a positive experience.' (P1023)

'The first time I defeated the boss The Nameless King was an overwhelmingly positive experience.' (P1236)

'It's a phenomenal experience succeeding at fighting enemies I previously struggled with, using a much more powerful character.' (P1510)

Theme	*Overall (N = 95)*	*Positive (n = 57)*	*Negative (n = 38)*	χ^2	*p*
Achievements & Victories	58 (61%)	**45 (79%)**	**13 (34%)**	**17.355**	**< .001**
Learning & Improvement	43 (45%)	**37 (65%)**	**6 (16%)**	**20.268**	**< .001**
Difficulties & Failures	78 (82%)	49 (86%)	29 (76%)	0.863	.353
Enemy Encounters	64 (67%)	43 (75%)	21 (55%)	3.354	.067

Table 1. The overall absolute and relative frequencies of the identified themes are depicted in the second column. In the third and fourth column these frequencies are split by valence of the experience (positive and negative) and tested for statistical significant differences with Pearson's Chi-squared tests with Yates' continuity correction. Themes with statistical significant differences are shown in bold.

The overall valence was coded negative, if the play session was mainly described with negative outcomes such as frustrations, anger or general negativity:

'I was hit by a special attack and died right before I was able to kill them. It was a terribly negative experience.' (P1410)

'Ultimately the feelings boiled down to excitement, anger, and disappointment.' (P643)

Measures

To investigate the role of affect, competence, challenge and challenge-skill balance, several scales were included. All measures employed 7-point Likert scales ranging from strongly disagree (1) to strongly agree (7).

Positive and negative affect

The general affective quality of the experiences was assessed with the International Positive and Negative Affect Schedule Short Form (I-PANAS-SF; [40]). It is widely used to measure strong positive ($M = 5.51$, $SD = 1.11$, Cronbach's $\alpha = .74$) and negative ($M = 3.13$, $SD = 1.18$, Cronbach's $\alpha = .60$) affective states.

Player Experience Need Satisfaction

The Player Experience Need Satisfaction scale (PENS; [34]), was used to examine the role of autonomy ($M = 5.65$, $SD = 1.23$, Cronbach's $\alpha = .83$), competence ($M = 5.42$, $SD = 1.16$, Cronbach's $\alpha = .71$), and relatedness ($M = 4.19$, $SD = 1.49$, Cronbach's $\alpha = .74$) with three items each, as already done in previous PX research (e.g., [5]).

Challenge

Perceived challenge of the situation ($M = 5.85$, $SD = 1.15$, Cronbach's $\alpha = .85$) was measured with the 5 items for challenge adapted from the Game Experience Questionnaire [16]. This subscale included items such as *'I thought it was hard'* and *'I had to put a lot of effort into it'*.

Challenge-Skill Balance

Challenge-Skill Balance ($M = 5.72$, $SD = 1.03$, Cronbach's $\alpha = .62$) was measured as agreement on the 3-item scale developed by Schmierbach et al. [35]: *'I was challenged, but I believed my skills would allow me to meet the challenge'*, *'My abilities matched the challenge of the situation'*, and *'The game kept me on my toes but did not overwhelm me.'*

Enjoyment

Enjoyment ($M = 6.56$, $SD = 0.98$, Cronbach's $\alpha = .95$) was measured using a 3-item scale, originally developed and validated by Oliver and Bartsch [28]. Participants indicated agreement on the following statements *'It was fun for me to play'*, *'I had a good time playing'*, and *'The experience was entertaining'*.

RESULTS

Overall frequencies of themes and overall valence

Taken together, the identified themes *achievements & victories, learning & improvement, difficulties & failures, lack of progress* and *enemy encounters* made up 87.8% of all reports, thus covering the most substantial part of all experiences. The overall frequency of each theme is depicted in Table 1. Out of the 95 experiences, 57 (60%) were coded as overall positive and 38 (40%) as an overall negative experience. This coding was used in the subsequent analysis to investigate characteristics of positive in comparison to negative overall experiences. The identified themes are described in the following section.

Themes of experience reports

The coding of open-ended answers identified the following themes, which hereby will be described in detail with the help of excerpts from player reports.

Achievements & Victories

Moments of *achievements & victories* were reported by 61% of all participants. Players typically reported these moments after defeating a boss-enemy or after a standard enemy-encounter: *'I slowly whittled the life from this evil force until I was victorious. A great feeling of joy and accomplishment washed over me and a sigh of relief left my lips.'* (P532)

Very rarely participants reported an achievement outside the context of a fight: *'I finally got the dragon head and torso stones for the first time in a souls-game [...] There is no joy like the satisfaction of finally getting to do something that you've waited to do for years.'* (P1293)

In the vast majority of all player reports about moments of *achievements & victories*, participants described their success in contrast to the high challenges imposed by the game, the unpredictability of outcomes and previous deaths and consequences experienced within their play session.

'I was so excited and so happy... The feeling you get after some tries, when you start to think something is just impossible and then you get through a difficult part of the game... That feeling is indescribable.' (P521)

'I was very close to death but I managed to defeat him as he was in the animation of swinging his weapon, which would have certainly killed me.' (P1236)

	Positive valence ($n = 57$)				Negative valence ($n = 38$)			
Scale	*Mean*	*SD*	*Median*	*Range*	*Mean*	*SD*	*Median*	*Range*
Positive affect[1]	5.93	0.75	6	4.2 - 7	4.88	1.28	4.8	1 - 7
Negative affect[1]	3.05	1.13	3	1 - 6	3.25	1.26	3	1.4 - 6.2
Competence[2]	5.79	0.77	6	4 - 7	4.87	1.41	5	1.67 - 7
Autonomy[2]	5.76	0.96	6	3 - 7	5.48	1.55	6	1 - 7
Relatedness[2]	4.45	1.33	4.67	1.33 - 7	3.81	1.58	6	1 - 7
Challenge[3]	6.17	0.76	6.2	3.8 - 7	5.38	1.46	6	1.8 - 7
Challenge-skill balance[4]	5.96	0.86	6	3.33 - 7	5.35	1.15	5.67	2.33 - 7
Enjoyment[5]	6.87	0.37	7	5 - 7	6.11	1.38	6.83	1 - 7

Table 2. Mean, standard deviation, median and range of player experience scales for experiences split by valence of the experience. Item sources: [1]I-PANAS-SF [40], [2]PENS [34], [3]GEQ [16], [4]Schmierbach et al. [35] and [5]Oliver and Bartsch [28].

Another remarkable characteristic of *achievements & victories* was their depiction in the light of fear and anxiety:

'*So far every souls-game makes me afraid of not knowing where to go and to know that you can lose your progress any time. This is especially pointed out in Dark Souls III. It makes you so angry, whenever you die from a boss after trying to beat him for like 20 minutes, but afterwards you just go for it and eventually complete it and that is the best feeling in the game!*' (P1141)

'*After several hours of trying, I finally beat the Soul of Cinder, and it was absolutely the best feeling I ever had in Dark Souls III. It was a big relief, I felt anxious and scared at the same time, but in a positive way. I mean it's as if you're accomplishing something big, like getting a degree.*' (P870)

'*I was afraid and not knowing if I will survive because of this new mechanic where an enemy evolves into a horrific being, which is something I would never have expected. But I felt proud of my feats, because they were against seemingly impossible odds.*' (P532)

Learning & Improvement

Moments of *learning & improvement* were reported by 45% of all participants. These reports typically contained narrations of figuring out certain techniques, strategies or combat patterns in order to progress within the game. For example: *'To have learnt the basics of parrying in such a short time frame made me feel really good about myself.'* (P1160)

'*I learned the right timing to evade his attacks, the right time frame to bring in a few hits myself and when to step back and heal.*' (P1060)

Moreover, moments of *learning & improvement* were typically evoked by avatar death, challenging moments and difficulties in general.

'*Obviously I died a lot, but every time I learned something new.*' (P1060)

'*My best experience yet was when I fought the Dancer of the Boreal Valley. It took me a total of 6 hours attempting to beat her [...] once I beat her, there has been no better feeling of satisfaction than seeing her hit the ground [...] I refused to watch strategy videos and learned all of her mechanics alone.*' (P1182)

Difficulties & Failures

Moments of *difficulties & failures* were reported by 82% of all participants, thus making this theme the most frequently reported. Narrations of *difficulties & failures* typically contained occurrences of avatar death, failed attempts to beat a boss, struggle from difficult gameplay and coping with high challenges.

'*We died to Yhorm (boss-enemy) since neither of us knew about the storm-ruler sword.*' (P1425)

'*I came up to the Dancer of the Boreal Valley boss, and I spent half an hour figuring out its spinning and the one-hit-kill grab attack. That horrible thing sighs in a peculiar way, and then lunges forward with a horizontal grab. If the grab connects, she picks me up and sticks her sword down through me.*' (P1164)

As previously depicted, *difficulties & failures* were often reported in context of *achievements & victories* and moments of *learning & improvement*.

Enemy Encounters

Enemy Encounters were reported by 67% of all participants. These reports typically contained narrations about a boss fight or a regular enemy in the game.

'*So I am off killing Nazguls left and right and eventually make my way to the first boss.*' (P54)

'*I was engaged in combat with a hollowed soldier, equipped with a spear and a metal greatshield.*' (P194)

'*I found a pretty hard enemy, which I didn't feel like fighting.*' (P254)

As previously depicted in the sections *Achievements & Victories, Learning & Improvement* and *Difficulties & Failures*, reports of *enemy encounters* often led to challenging gameplay moments, eventually leading to victory or failure of the player.

Interrelation of themes

Some of the participants' reports clearly showed a strong interrelation among all themes:

'The process most of the time looks like this: Feeling pumped for having reached a new boss fight, getting angry because that boss is impossible to defeat, starting to learn the bosses attacks and routines, getting nervous and nearly having a panic attack when getting the boss' health bar below 50%, defeating the boss, feeling wrecked but also happy for having accomplished something truly challenging.' (P751)

This narration of a participant demonstrates, how *achievements & victories, moments of learning & improvement*, *difficulties & failures* and *enemy encounters* are intertwined. However, there were a few experiences that did not fit in these categories. For example, this player described his or her admiration of the game world: *'Opening the doors to the High Wall of Lothric took my breath away, knowing that you could go practically anywhere from here. I stood there and just looked at the scenery for 5 minutes.'* (P1050)

Or a different player who felt lost in the game world, but also fascinated and inspired at the same time: *'But my first experience can only be described as finally arriving to a far off land, searching for my purpose. A sense of awe and fear stuck with me through my time with this game. I felt enthralled and on the edge constantly, never taking a second for a break on my quest.'* (P70)

The interrelation and frequency of themes split by valence will be reported in the next section.

Themes in positive and negative experiences

The frequency of each theme is shown split by valence in Table 1 to describe which themes are associated with a positive valence. Results of the Pearson Chi-squared tests with Yates' continuity correction show significantly more observations of *achievements & victories* ($\chi^2 = 17.36$, $p < .001$) and *learning & improvement* ($\chi^2 = 20.27$, $p < .001$) in positive than in negative experiences. However, *difficulties & failures* and *enemy encounters* did not occur significantly more often in positive than in negative experiences (see Table 1).

Achievements & Victories and Difficulties & Failures

A more detailed analysis showed that *difficulties & failures* occurred significantly more often (55 of 58; 95%) in experiences of *achievements & victories* than in experiences that did not report *achievements & victories* (23 of 37; 62%), $\chi^2_{df=1} = 14.26$, $p < .001$.

Learning & Improvement and Difficulties & Failures

The theme *difficulties & failures* occurred significantly more often (41 of 43; 95%) in experiences of *learning & improvement* than in experiences that did not report *learning & improvement* (37 of 52; 71%), $\chi^2_{df=1} = 7.8$, $p < .01$.

Analysis of player experience measures

To investigate whether the difference between overall positive and negative experience reports was mirrored in the quantitative measures, the data was split by valence. Descriptive statistics of these two groups are shown in Table 2. To increase the robustness of the results, the groups were compared on an ordinal scale with Mann-Whitney U tests.

Positive and Negative Affect

Positive affect was greater in overall positive experiences ($Mdn = 6$) than in overall negative experiences ($Mdn = 4.8$), $Z = 4.30$, $p < .001$, $r = .44$. This reflects the qualitative analysis, but also shows that experiences with a negative valence were still accompanied with relatively high positive affect. However, no significant differences in negative affect were found between positive ($Mdn = 3$) and negative experiences ($Mdn = 3$). This shows that even when players did report a negative outcome in the narrative, such as when they were constantly failing in a fight against a boss, these experiences were still rated as positive.

Player Experience Need Satisfaction

Competence need satisfaction was greater in overall positive experiences ($Mdn = 6$) than in overall negative experiences ($Mdn = 5$, $Z = 3.24$, $p < 0.01$, $r = .33$), for relatedness, the differences between positive ($Mdn = 4.7$) and negative (3.7) experiences was marginally significant, $Z = 1.95$, $p = .052$, $r = 0.20$. For autonomy, no significant difference between the groups (both $Mdn = 6$) was observed.

Challenge-skill balance

Data showed that challenge-skill balance was greater in overall positive experiences ($Mdn = 6$) than in overall negative experiences ($Mdn = 5.7$), $Z = 2.53$, $p < .05$, $r = .26$. However, in both groups the perceived balance was very high, showing that although many players reported difficulties and failures, the challenge of the game was still perceived as very well matched with their skill.

Challenge

Investigating the level of perceived challenge isolated from challenge-skill balance, data showed that challenge was greater in overall positive experiences ($Mdn = 6.2$) than in overall negative experiences ($Mdn = 6$), $Z = 2.56$, $p < .05$, $r = .26$. Again, in both groups the last game-session was perceived as highly challenging, with only a marginal difference between them.

Enjoyment

Enjoyment was very high for overall positive experiences ($Mdn = 7$) and overall negative experiences ($Mdn = 6.8$), $Z = 3.75$, $p < .001$, $r = .38$. Although this difference is statistically significant, it shows that even negative experiences in the end were perceived as enjoyable.

Prediction of overall valence

A binomial logistic regression was used to identify important predictors for the valence of the experience among qualitative (the 4 themes) and quantitative variables, as well as to gain an understanding of their relative importance. As depicted in Table 3, results show that *achievements & victories* and *learning & improvement* are the only significant predictors in the model. The odds ratio for these predictors suggest that if *achievements & victories* were reported, the chance was almost 4 times higher that a positive experience occurred than for those experiences that did not report such an event. For *learning & improvement* the relative chance even increased to more than 5 times. Additionally, the results show that neither the PENS subscales, nor measures of challenge or

	Predictor	β	SE β	Wald's χ^2	df	p	e^{β} (odds ratio)	95% CI of e^{β}	
Qualitative	Constant	–0.454	0.619	36.00	1	.85	0.635	0.185	2.182
	Achievements & Victories	1.345	0.685	**3.90**	**1**	**<.05**	3.838	1.025	15.602
	Learning & Improvement	1.615	0.678	**5.70**	**1**	**<.05**	5.030	1.365	20.212
	Difficulties & Failures	-1.064	0.780	1.90	1	.17	0.345	0.070	1.541
	Enemy Encounters	0.486	0.665	0.53	1	.46	1.626	0.435	6.067
Quantitative	Competence	0.589	0.359	2.70	1	.10	1.802	0.918	3.801
	Autonomy	-0.185	0.333	0.31	1	.58	0.831	0.424	1.586
	Relatedness	0.276	0.338	0.67	1	.41	1.318	0.686	2.624
	Challenge-skill balance	0.310	0.327	0.90	1	.34	1.364	0.722	2.633
	Challenge	0.521	0.387	1.80	1	.18	1.683	0.815	3.774

Model fit tests	χ^2	df	p
Overall model evaluation			
Likelihood ratio test	45.34	9	<.001
Score test	38.68	9	<.001
Goodness-of-fit test			
Hosmer & Lemeshow	7.43	8	.49

Table 3. Statistics of the binomial logistic regression predicting the valence of the experience (0 = negative, 1 = positive) using the 4 identified themes as nominal predictors and the 5 player experience scales as interval scaled predictors. For each predictor, standardized regression coefficients β, the corresponding standard errors SE β and the results of a Wald χ^2 test with the degrees of freedom and the corresponding p-value are shown. The Wald tests show that the regression coefficients of the themes *Achievements & Victories* and *Learning & Improvement* are significantly different from 0. e^{β} depicts the odds ratio of the predictor, i.e. the relative chance of having an experience with a positive valence when the theme appeared in the narrative for qualitative predictors, or, for quantitative predictors, when the rating on the scale increased by 1. The *likelihood ratio test* and the *score test* show a significant improvement of the model fit compared to the null model. The *Hosmer & Lemeshow* test investigates the null hypothesis that observed rates of positive and negative valence is equal for all subgroups of predicted probabilities. Additional model fit statistics and pseudo R^2: Nagelkerke R^2 (Max rescaled R^2) = .513. Kendall's $Tau-\alpha$ = .355. Goodman-Kruskal Gamma = .734. Somers's D_{xy} = .733. c-statistic = 86.6%. Model performance: Accuracy = .82, Precision = .83, Negative predictive value = .80, Sensitivity = .88, Specificity = .74.

challenge-skill balance were significantly predictive for the overall valence of the reported experience.

DISCUSSION

The aim of this study was to examine factors contributing to a positive experience in a game that is defined by high challenges and countless, punishing avatar deaths. We found that most players rated their experience with *Dark Souls III* as enjoyable, as reflected in the high ratings on enjoyment, positive affect and numerous positive experience reports. A higher challenge was more likely to be associated with positive than with negative experiences, as shown by heightened subjective challenge scores. Reports on *difficulties & failures* however did not significantly differ between the groups, suggesting that they are not an exclusive characteristic of either positive or negative experiences. Positive affect, competence and relatedness were associated with positive reports, whereas negative affect and autonomy did not differ between positive and negative experiences. Furthermore, moments of *achievements & victories* and *learning & improvement* occurred significantly more often in positive experiences. Moreover, these two themes were two important predictors for positive reports, *learning & improvement* being the overall strongest predictor. The vast majority of players in our study reported numerous negative events, which typically have to be avoided in order to progress, *difficulties & failures* being the most frequently mentioned negative event. Avatar death and high challenges thus are a substantial part of the game. As previously mentioned, these seemingly negative events occurred equally often in negative and positive experiences, indicating that even if players experienced *difficulties & failures*, their play session could still be an overall positive experience. Whether players enjoy the game despite facing negative events or if negative events bear a potential to form a positive experience, is discussed in the following section.

Learning & Improvement

The first contributor and strongest predictor of a positive experience was *learning & improvement*, supporting the notion that learning makes video games "fun" [26], especially if they provide a lot of complex information and patterns to understand and figure out. A closer look at how *learning & improvement* emerged in *Dark Souls III* further supports Kosters' [26] notion of learning: Moments of *learning & improvement* typically consisted of situations where players were figuring out new information, such as enemy patterns or where they were improving on their performance after a string of failures, also coinciding with Juul [18], stating that avatar death is the death you survive and learn from. In total, 95% of all participants who reported moments of *learning & improvement* also reported *difficulties & failures*. It therefore seems that a challenging gameplay, where players are likely to fail, may enable learning processes, which eventually lead to a performance improvement. These findings coincide with Iacovides et al. [15] who showed that breakdowns (caused by challenges) do lead to learning if there was a breakthrough in understanding how to solve the problem in the game. The present study extends our understanding of learning moments in video games by showcasing their accentuated role in a context of high challenges and numerous avatar deaths. In a game, where *difficulties & failures* are frequent, moments of learning are especially important for a positive experience. As seen in player reports, moments of learning to a large extent stem from *difficulties & failure*, further emphasizing the role of high challenges and avatar death as a learning mechanic.

Achievements & Victories

In line with previous player experience research (e.g. [5]), the second important contributor to positive experiences identified in this study was *achievements & victories*. A closer look at how achievements in *Dark Souls III* are characterized and described sheds some light on their role in a high challenge context: Narrations of *achievements & victories* usually depicted a victorious boss fight or another challenging enemy encounter, typically after a series of failed attempts. Similar to previous work on permadeath in *DayZ* [2, 7], where high risks and consequences increased players' sense of involvement and meaning, players in the present study often rated their achievement as particularly satisfying in view of previous failed attempts and struggles. Player reports of victory and success clearly depict that, much like in *DayZ*, grave consequences resulting from avatar death and high challenges in *Dark Souls III* form a general atmosphere of anxiety, fear of imminent loss and the player's awareness of everything being at stake. This is in line with the general assumption that excessive challenge levels beyond the challenge-skill balance lead to anxiety [11]. However, anxiety itself in this case emphasized the achievement as a hard-earned success amidst unpredictable outcomes and high challenges. This might be an indication that a positive experience and enjoyment are not exclusively tied to optimal challenge-skill balance and flow. It furthermore raises questions about the significance of these concepts and if they are adequate for every type of player experience, since excessive difficulty in the present case was able to form meaningful achievements. Yet, since the game appeals to a rather specific audience, which will be addressed in the Limitations section, it is difficult to draw concrete conclusions concerning this matter based on the results of this study.

Considering players' accounts of successful moments, it comes as little surprise that 95% of all participants who reported *achievements & victories* also reported narrations of difficulties and failure. Furthermore, when looking only at participants who experienced *difficulties & failure*, *achievements & victories* occurred significantly more often in the positive experience group. As a contribution to player experience research, we showcased the interplay of positive and negative experiences, emphasizing that not only are they not mutually exclusive, but one actually to a large extent depends on the other. Applying a different methodological approach than Allison et al. [2], our results confirm that the meaning of achievements is heightened through severe consequences from avatar death. In addition, the present study was able to showcase this finding not only with qualitative reports but analyzed in more detail frequencies of the themes *achievements & victories* in relation to *difficulties & failures*. Achievements themselves therefore may be directly linked to positive emotions [5], but in *Dark Souls III* achievements to a large extent are characterized by avatar death and high challenges. Whereas Bopp et al. [5] demonstrated that intense, negative emotions such as sadness directly contribute to positive player experiences, our results show that negative events such as avatar death and difficulties do not directly predict a positive experience, but they enable and characterize moments of achievement and learning.

Challenge and Enjoyment

In contrast to studies conducted by Schmierbach et al. [35], Klarkowski et al. [23] and Gutwin et al. [13], which showed lower enjoyment scores for higher difficulties, a higher challenge in the present study was more likely to be associated with a positive experience. Moreover, enjoyment and positive affect ratings, while being significantly lower in the negative experience group, were generally very high even in the negative group, showing that negative experiences did not lead to low enjoyment. This might be due to the players' awareness that high challenges and avatar death are a crucial component for the enjoyment of *Dark Souls III*. Thus, a play session with negative experiences might still be perceived as enjoyable, if players realize that they are in a process of learning and thus perceive failing as part of a bigger experience. Our results indicate that difficult games are not in general less enjoyable as assumed by Schmierbach et al. [35]. Not only did *difficulties & failures* occur as often in positive as in negative experiences, but they for many players enabled and formed *achievements & victories* and moments of *learning & improvement* for positive experiences, thus demonstrating, how close negative and positive events were intertwined in *Dark Souls III*.

In conclusion, the present study emphasizes the role of *achievements & victories* and *learning & improvement* in a highly challenging context in order to reach a positive experience. Achievements themselves are perceived as outstanding experiences in contrast to the state of anxiety and fear the game creates through its excessive challenges and impactful avatar death mechanic. Victories over enemies were enjoyed so much because players had to earn them the hard way. This finding goes in line with Przybylski et al. [32], showing that rewards provide competence need-satisfaction and increase motivation if the game offers a challenging gameplay. Reward may thus have even more value, if the game contains higher challenges. Hard-earned achievements seem to make up for frustrations and negative moments in the game, a role which may be not as emphasized in less challenging games. Although *Dark Souls III* is not a casual game, these implications may also be of interest for a broader market. The implementation of difficult-to-reach goals adds a further challenge and depth to a game, committing players who are willing to devote more time to reach out for more difficult goals. In casual games the achievement of difficult goals may be not as critical for an enjoyable experience but they could nevertheless add an additional layer of depth and appeal for some players. High challenges and punishing avatar death mechanics may not be crucial to all type of games for a positive experience, but they seem to play a substantial role in enabling learning processes, thus making achievements seem hard-earned and meaningful.

Limitations

To explore why some players enjoy video games with excessive difficulty, we specifically recruited participants with the help of online fan forums of *Dark Souls*. We felt that this was a necessary step, since the game has a very steep learning curve at the beginning and to have novices play the game would have been a completely different (albeit interesting) study. This procedure however most likely led to a very specific sample who might have had a strong positive bias towards this kind of

games. This becomes evident when taking a closer look at the sample of this study: Participants were mainly male players (95.8%) with an average age of less than 24 years. Almost all participants reported experience with other games of this series (e.g. *Dark Souls* or *Dark Souls II*) and therefore might have a strong affinity for challenging games, respectively a tolerance for high difficulty gameplay. The participants of this study also probably knew what to expect from *Dark Souls III* and their skill level to a certain degree most likely matched the high demands of the game, which made it possible to reach a positive experience in the first place. Hence, the results of this study may in some respects be dependent on personal preferences and player personalities and can therefore not be generalized for all types of players. As stated by Yun et al. [44], when discussing game enjoyment it is not sufficient to merely categorize players according to their level of experience in video games, another basic criterion is if a players' primary objective in a game is seeking challenges, victories or a balanced version of both. Although we found moments of *achievements & victories* and *learning & improvement* to be crucial elements of positive experiences, one could argue that these factors are especially important to players specifically seeking challenges (*challenge seekers*), but may not play a predominant role for other player personalities. Juuls notion of players' repertoire [21] applies in this context in two different ways. First, challenge in *Dark Souls III* is based on learning how to control the character and on reading enemy attack and movement patterns, i.e. the player needs to build up a repertoire of skills and strategies throughout the game and adapt and refine them as he or she encounters new enemies. Second, the perception of difficulty is shaped by the players' experience with previous games that they have played and the skills and strategies they have acquired. Since most participants of this study were already experienced with the *Dark Souls* series, they most likely built up a repertoire of skills and strategies that helped them to reach the game's goals, thus influencing the perception of the difficulty in *Dark Souls III* and possibly other games as well. It would therefore be up to future research to explore in detail, how different player personalities with different skill repertoires influence the meaning of *achievements & victories* and *learning & improvement* in video games.

The results may also not be generalized for all game genres and designs, since this study merely discusses one specific type of game. The action oriented gameplay of *Dark Souls III* focuses heavily on real-time melee combat and therefore offers a very specific type of challenge. The game does not cover all aspects of gameplay difficulty and even within the RPG-genre a game can be challenging because of various reasons, such as hard to solve puzzles or complex turn-based combat systems. However, *Dark Souls III* still features many common sources of difficulty such as learning different enemy patterns, developing new strategies, character control, reaction speed and luck.

The present exploratory study identified *achievements & victories* and *learning & improvement* as predominant factors associated with positive experiences in a setting of high challenges. No conclusions can be made concerning the causal effect of high-challenge settings on positive experiences if, for example, compared to low-challenge settings. It would be up to future work, to compare different game difficulty levels with regard to achievement and learning experiences. Still, the present study identified some important predictors for positive experiences and their relationship to negative events, which should be taken into account, when discussing the role of challenge and avatar death in any video game.

Furthermore, participants in the present study were asked to report an outstanding experience from their memory and may have missed out or altered important information, which in hindsight were perceived differently. For instance, the perceived challenge-skill balance is associated with a positive experience, which coincides with Klarkowski et al. [23], demonstrating higher enjoyment scores for balanced experiences. Considering the numerous occurrences of *difficulties & failures* across all experiences in our study, it is unclear though, how strong the perceived challenge-skill balance in the end may be influenced by recent achievements in the game. Gutwin et al. [13] demonstrated an effect on players' perception focusing on peaks and ends of the game. It is therefore a reasonable assumption that a play session in hindsight may seem more balanced after a glorious victory than it actually was, thus relativizing the role of challenge-skill balance as it is currently discussed in research (e.g. [13, 23]). As the comparison of negative events and high enjoyment ratings showed, qualitative and quantitative measures may highlight different aspects of an experience and therefore a combination of both allows for a more nuanced discussion. Future research applying subjective and behavioral measures may provide further insights concerning actual events in the game and how they were perceived by players.

Conclusion

An appropriate challenge-skill balance is usually seen as being substantial for positive player experiences. However, the present study found that negative events such as *difficulties & failures* characterize moments of *achievements & victories* and *learning & improvement*. These moments to a large extent explain positive player experiences in a game with excessive difficulty. Players' reports indicate that high challenges and avatar death make moments of achievement meaningful and thus, enable positive player experiences.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their thoughtful comments and Lena Aeschbach for assisting with the literature research. Please note that the first and the second author were the main contributors to this study.

REFERENCES

1. Justin T. Alexander, John Sear, and Andreas Oikonomou. 2013. An investigation of the effects of game difficulty on player enjoyment. *Entertainment Computing* 4, 1 (2013), 53–62. http://www.sciencedirect.com/science/article/pii/S1875952112000134

2. Fraser Allison, Marcus Carter, and Martin Gibbs. 2015. Good Frustrations: The Paradoxical Pleasure of Fearing

Death in DayZ. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. ACM, 119–123. http://dl.acm.org/citation.cfm?id=2838810

3. Maria-Virginia Aponte, Guillaume Levieux, and Stéphane Natkin. 2011. Difficulty in videogames: an experimental validation of a formal definition. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. ACM, 49. http://dl.acm.org/citation.cfm?id=2071484
4. Scott Bateman, Regan L. Mandryk, Tadeusz Stach, and Carl Gutwin. 2011. Target assistance for subtly balancing competitive play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2355–2364. http://dl.acm.org/citation.cfm?id=1979287
5. Julia Ayumi Bopp, Elisa D. Mekler, and Klaus Opwis. 2016. Negative Emotion, Positive Experience? Emotionally Moving Moments in Digital Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2996–3006. http://dl.acm.org/citation.cfm?id=2858227
6. Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa
7. Marcus Carter, Martin Gibbs, and Greg Wadley. 2013. Death and dying in DayZ. In *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*. ACM, 22. http://dl.acm.org/citation.cfm?id=2513013
8. Jared E. Cechanowicz, Carl Gutwin, Scott Bateman, Regan Mandryk, and Ian Stavness. 2014. Improving player balancing in racing games. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*. ACM, 47–56. http://dl.acm.org/citation.cfm?id=2658701
9. Jenova Chen. 2007. Flow in games (and everything else). *Commun. ACM* 50, 4 (2007), 31–34. http://dl.acm.org/citation.cfm?id=1232769
10. Amra Copcic, Sophie McKenzie, and Michael Hobbs. 2013. Permadeath: A review of literature. In *2013 IEEE International Games Innovation Conference (IGIC)*. IEEE, 40–47. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6659156
11. M. Csikszentmihalyi. 1990. *Flow. The Psychology of Optimal Experience.* New York: Harper Perennial.
12. Emily Flynn-Jones. 2015. Don't Forget to Die: A Software Update is Available for the Death Drive. In *The Dark Side of Game Play: Controversial Issues in Playful Environments*, T. Mortensen, J. Linderoth, and A. Brown (Eds.). London: Routledge, 50–66.
13. Carl Gutwin, Christianne Rooke, Andy Cockburn, Regan L. Mandryk, and Benjamin Lafreniere. 2016. Peak-End Effects on Player Experience in Casual Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5608–5619. http://dl.acm.org/citation.cfm?id=2858419
14. Robin Hunicke. 2005. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*. ACM, 429–433. http://dl.acm.org/citation.cfm?id=1178573
15. Ioanna Iacovides, Anna L. Cox, Patrick McAndrew, James Aczel, and Eileen Scanlon. 2015. Game-play breakdowns and breakthroughs: exploring the relationship between action, understanding, and involvement. *Human–computer interaction* 30, 3-4 (2015), 202–231. http://www.tandfonline.com/doi/abs/10.1080/07370024.2014.987347
16. Wijnand IJsselsteijn, Wouter Van Den Hoogen, Christoph Klimmt, Yvonne De Kort, Craig Lindley, Klaus Mathiak, Karolien Poels, Niklas Ravaja, Marko Turpeinen, and Peter Vorderer. 2008. Measuring the experience of digital game enjoyment. In *Proceedings of Measuring Behavior*. Noldus Information Technology Wageningen, Netherlands, 88–89.
17. Seung-A. Annie Jin. 2012. Toward integrative models of flow: Effects of performance, skill, challenge, playfulness, and presence on flow in video games. *Journal of Broadcasting & Electronic Media* 56, 2 (2012), 169–186. http://www.tandfonline.com/doi/abs/10.1080/08838151.2012.678516
18. Jesper Juul. 1999. *A clash between game and narrative*. Master's thesis. University of Copenhagen. http://www.jesperjuul.net/thesis/AClashBetweenGameAndNarrative.pdf
19. Jesper Juul. 2009. Fear of failing? The many meanings of difficulty in video games. In *The Video Game Theory Reader*, Mark J. P. Wolf and Bernard Perron (Eds.), Vol. 2. New York: Routledge, 237–252.
20. Jesper Juul. 2010. *A casual revolution: Reinventing video games and their players*. Cambridge, Mass: MIT Press.
21. Jesper Juul. 2011. *Half-real: Video games between real rules and fictional worlds*. Cambridge, Mass: MIT Press.
22. Jesper Juul. 2013. *The art of failure: An essay on the pain of playing video games*. Cambridge, Mass: MIT Press.
23. Madison Klarkowski, Daniel Johnson, Peta Wyeth, Mitchell McEwan, Cody Phillips, and Simon Smith. 2016. Operationalising and Evaluating Sub-Optimal and Optimal Play Experiences through Challenge-Skill Manipulation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5583–5594. http://dl.acm.org/citation.cfm?id=2858563

24. Lisbeth Klastrup. 2007. Telling & sharing? Understanding mobile stories & future of narratives. In *7th international digital arts and culture conference the future of digital media culture*. Perth, Australia. http://www.leonardo.info/LEA/perthDAC/LKlastrup_LEA160203.pdf

25. Christoph Klimmt, Christopher Blake, Dorothée Hefner, Peter Vorderer, and Christian Roth. 2009. Player performance, satisfaction, and video game enjoyment. In *International Conference on Entertainment Computing*. Springer, 1–12. http://link.springer.com/chapter/10.1007/978-3-642-04052-8_1

26. Raph Koster. 2013. *Theory of fun for game design*. Sebastopol, CA: O'Reilly Media.

27. Nicole Lazzaro. 2004. Why we play games: Four keys to more emotion without story. Presentation at the Game Developers Conference. http://www.xeodesign.com/whyweplaygames/xeodesign_whyweplaygames.pdf. (2004). Accessed: 2016-09-19.

28. Mary Beth Oliver and Anne Bartsch. 2010. Appreciation as Audience Response: Exploring Entertainment Gratifications Beyond Hedonism. *Human Communication Research* 36, 1 (Jan. 2010), 53–81. DOI: http://dx.doi.org/10.1111/j.1468-2958.2009.01368.x

29. 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 (2012), 175–196. http://www.tandfonline.com/doi/abs/10.1080/15213269.2012.673850

30. Felipe Pepe. 2016. The history of the Quest Compass & its dreadful convenience. http://www.gamasutra.com/blogs/FelipePepe/20160412/270100/The_history_of_the_Quest_Compass__its_dreadful_convenience.php. (April 2016). Accessed: 2016-09-21.

31. Helmut Prendinger, Kamthorn Puntumapon, and Marconi Madruga. 2016. Extending Real-Time Challenge Balancing to Multiplayer Games: A Study on Eco-Driving. *IEEE Transactions on Computational Intelligence and AI in Games* 8, 1 (2016), 27–32. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6932459

32. Andrew K. Przybylski, C. Scott Rigby, and Richard M. Ryan. 2010. A motivational model of video game engagement. *Review of general psychology* 14, 2 (2010), 154. http://psycnet.apa.org/journals/gpr/14/2/154/

33. Ute Ritterfeld, Michael Cody, and Peter Vorderer. 2009. *Serious games: Mechanisms and effects*. New York: Routledge.

34. Richard M. Ryan, C. Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* 30, 4 (2006), 344–360. http://link.springer.com/article/10.1007/s11031-006-9051-8

35. Mike Schmierbach, Mun-Young Chung, Mu Wu, and Keunyeong Kim. 2014. No One Likes to Lose. *Journal of Media Psychology* 26, 3 (2014), 105–110. http://dx.doi.org/10.1027/1864-1105/a000120

36. John L. Sherry. 2004. Flow and media enjoyment. *Communication theory* 14, 4 (2004), 328–347. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2885.2004.tb00318.x/abstract

37. Jan D. Smeddinck, Regan L. Mandryk, Max V. Birk, Kathrin M. Gerling, Dietrich Barsilowski, and Rainer Malaka. 2016. How to Present Game Difficulty Choices? Exploring the Impact on Player Experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5595–5607. http://doi.acm.org/10.1145/2858036.2858574

38. Penelope Sweetser and Peta Wyeth. 2005. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)* 3, 3 (2005), 3–3. http://dl.acm.org/citation.cfm?id=1077253

39. Chin Hiong Tan, Kay Chen Tan, and Arthur Tay. 2011. Dynamic game difficulty scaling using adaptive behavior-based AI. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 4 (2011), 289–301. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5783334

40. Edmund R. Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology* 38, 2 (2007), 227–242. http://jcc.sagepub.com/content/38/2/227.short

41. Rodrigo Vicencio-Moreira, Regan L. Mandryk, and Carl Gutwin. 2015. Now you can compete with anyone: Balancing players of different skill levels in a first-person shooter game. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2255–2264. http://dl.acm.org/citation.cfm?id=2702242

42. Peter Vorderer, Christoph Klimmt, and Ute Ritterfeld. 2004. Enjoyment: At the heart of media entertainment. *Communication theory* 14, 4 (2004), 388–408. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2885.2004.tb00321.x/abstract

43. Douglas Wilson and Miguel Sicart. 2010. Now it's personal: on abusive game design. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology*. ACM, 40–47. http://dl.acm.org/citation.cfm?id=1920785

44. Chang Yun, Philip Trevino, William Holtkamp, and Zhigang Deng. 2010. PADS: enhancing gaming experience using profile-based adaptive difficulty system. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*. ACM, 31–36. http://dl.acm.org/citation.cfm?id=1836140

Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples

Florian Brühlmann[1], Serge Petralito, Lena F. Aeschbach, and Klaus Opwis

Center for Cognitive Psychology and Methodology

University of Basel

Missionsstrasse 62a

CH-4055 Basel, Switzerland

Affiliation

[1] Corresponding author. Tel.: + 41 (0)61 207 06 66

Email addresses: florian.bruehlmann@unibas.ch (Florian Brühlmann), s.petralito@unibas.ch (Serge Petralito), lena.aeschbach@unibas.ch (Lena F. Aeschbach), klaus.opwis@unibas.ch (Klaus Opwis)

Abstract

Research in various academic fields relies increasingly on online samples. With the advent of crowdsourcing platforms, online data collection has become more popular than ever, although concerns have been raised recently. These concerns regard the data quality of these samples and the possible adverse effects of poor data on experimental manipulations and scale properties. Presently, research on carelessness in crowdsourced surveys is scarce. Therefore, the goal of this study (N = 394) was to systematically identify careless and inattentive behavior in a crowdsourced sample by applying various measures and methods for detecting carelessness. Results revealed that approximately half of all participants were inattentive in the online survey. Furthermore, carelessness and inattentive behavior appear highly task-dependent, because correlations between open answer quality and other measures were rather low. Thus, based on a predictive model and ease of interpretation, we recommend assessing the data quality of crowdsourced samples with a self-reported single item, one or multiple attention checks (such as an Instructed Response Item (IRI)), a LongString analysis, and a task-specific measure. This combination of detection methods accurately predicted careless participants, and excluding these participants increased the effect size in an experiment included in the survey.

Keywords: Inattentive responding; Careless responding; Crowdsourcing; Response patterns; Open answer; Latent profile analysis

Half of the Participants in Online Surveys Respond Carelessly: An Investigation of Data Quality in Crowdsourced Samples

Introduction

Online surveys have become a standard method of data collection in various fields, such as in recent psychological research (Gosling & Mason, 2015) and market research (Comley, 2015). Whereas in 2003 and 2004 only 1.6% of articles published in APA journals used the Internet (Skitka & Sargis, 2006), Gosling and Mason (2015) stated just a few years later that "studies that use the Internet in one way or another have become so pervasive that reviewing them all would be impossible" (p. 879). Moreover, this method covers virtually all areas of psychology. Online data collection has numerous advantages over laboratory studies: lower infrastructure costs (no laboratory infrastructure or individual time slots are needed), faster and cheaper data collection (Casler, Bickel, & Hackett, 2013; de Winter, Kyriakidis, Dodou, & Happee, 2015), more extensive distribution of the study, and lower hurdles for participation (Kan & Drummey, 2018). One of the most popular recruitment methods for participants in online studies for psychological research is the use of crowdsourcing services, such as Amazon's Mechanical Turk (MTurk) or FigureEight (formerly known as CrowdFlower). Regarding MTurk, approximately 15'000 published articles used this crowdsourcing platform between 2006 and 2014 for their data collection (J. Chandler & Shapiro, 2016; Kan & Drummey, 2018). On these platforms, various small tasks are offered in exchange for money to "crowd workers". All the advantages of other online data collection methods, such as cost- and time-effectiveness, also apply to crowdsourcing platforms (Kan & Drummey, 2018). Additionally, crowdsourcing platforms offer a more diverse population compared to typically homogenous samples from psychological studies (Kan & Drummey, 2018): In the case of MTurk, these workers are composed of a demographic containing more than 500'000 individuals from 190 countries (Paolacci & Chandler, 2014). While concerns considering the generalizability and validity of crowdsourced online samples have been discussed (Kan & Drummey, 2018), Gosling and Mason (2015) also reported that the mean and range of ages from an MTurk-sample are more representative of the general US population than a sample merely consisting of undergraduate students. Moreover, in comparison to online samples

recruited on social media platforms, some crowdsourced samples were found to have a higher diversity in terms of age, cultural, and socioeconomic factors (Casler et al., 2013), and more balanced gender ratios (de Winter et al., 2015). Furthermore, Kan and Drummey (2018) stated that MTurk is a viable alternative to traditional methods of data collection, because many studies showed similar patterns of findings in their crowdsourced data when compared to results using traditional approaches of data collection (Kan & Drummey, 2018, p. 244). However, given the increased distance between researchers and participants in online studies, and the possible influence of distractions in an uncontrolled setting, data collected online may suffer from bad quality stemming from inattentiveness and other forms of deceptive behavior. Participant carelessness or inattentiveness (Meade & Craig, 2012), have recently received increased attention from various researchers regarding their reasons, effects, detection, and prevention (Maniaci & Rogge, 2014; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). Although carelessness may also occur in laboratory studies, the problem seems especially common within online samples because survey administrations are often unproctored (Cheung, Burns, Sinclair, & Sliter, 2017; Fleischer, Mead, & Huang, 2015). Maniaci and Rogge (2014) went further, claiming that the current "replication crisis" in psychology may be related to careless respondents who take part in online surveys with insufficient attention. Regarding crowdsourced samples, concerns about the data quality have also been raised, as these workers are usually non-naive participants with possibly deceptive behavior. This tendency is exacerbated by the incentive-structure of these platforms. Further, the responses are often conducted in uncontrolled and possibly distracting environments (J. Chandler, Mueller, & Paolacci, 2014; J. Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015; Kan & Drummey, 2018; Peer, Brandimarte, Samat, & Acquisti, 2017; Stewart, Chandler, & Paolacci, 2017).

Causes and effects of carelessness

In the present study, we primarily focus on participant inattention or careless response. Other forms of invalid responding and deceptive behavior (such as social desirability and faking responses), also decrease data quality, but may have different causes and effects (Maniaci &

Rogge, 2014; McKay, Garcia, Clapper, & Shultz, 2018). Participant inattention might have many sources, one of them being the anonymity of computer-based surveys, which can result in a lack of accountability (Douglas & McGarty, 2001; Lee, 2006; Meade & Craig, 2012). Further important factors affecting carelessness in survey data are respondent interest, length of survey, social contact, and environmental distraction (Meade & Craig, 2012). Extrinsic motivation might also account for carelessness, such as when participants are paid for their answers. Gadiraju, Kawase, Dietze, and Demartini (2015) found that some participants recruited via crowdsourcing services employ strategies to minimize their invested time or effort in return for participation compensation. In these cases, careless responding and subsequent poor data quality emerge from crowdworkers who are solely interested in receiving their payment as fast as possible without providing valid data for the researcher. Furthermore, Niessen et al. (2016) observed that students also strove to complete surveys as quickly as possible in exchange for course credits. Aside from these external factors, a study conducted by McKay et al. (2018) found that careless responding is strongly related to malevolent personality traits, whereas its connection to benevolent traits was less pronounced.

Base rate estimates for bad online data quality stem from different concepts of invalid responding and different sources for online data collection: Recent research has estimated that, depending on the method, between 10% to 12% of participants in an online survey exhibit an answering behavior described as insufficient effort responding or careless responding (Meade & Craig, 2012). In a more heterogeneous sample, Maniaci and Rogge (2014) found that between 3% to 9% of participants respond carelessly. In an online survey with students from a university in the United States, Ward, Meade, Allred, Pappalardo, and Stoughton (2017) showed that 23% of the participants were flagged by at least one IRI. Collecting online data from a Facebook sample, Dogan (2018) estimated careless responding in 40.7% to 59.8% of the sample, depending on the measure used to detect careless behavior. For a crowdsourced sample on MTurk, Kan and Drummey (2018) found that between 21.8% and 55.8% of the sample (depending on eligibility requirements), proved deceptive and provided false data. However, it is important to note that Kan and Drummey (2018) did not assess carelessness or inattentive behavior. They solely refer to deceptive behavior concerning screening

requirements on Amazon's MTurk, which emerge under certain eligibility constraints. Carelessness in Hauser and Schwarz (2016) was assessed merely by using an instructional manipulation check, resulting in highly volatile estimates from 4% to 74.5%, depending on the exact method used. Instructional manipulation checks have also been criticized for being too restrictive, as partially skipping instructions does not automatically mean that participants are inattentive (Maniaci & Rogge, 2014). In another study assessing carelessness in a crowdsourced sample, Peer et al. (2017) found that only 27% of all participants in a FigureEight-sample passed all attention checks, and approximately 18% failed in all of them. While these numbers provide some valuable insights for assessing careless behavior in crowdsourced samples, the study did not include other carelessness measures. Therefore, it only identified one behavioral form of inattention or carelessness. Consequently, these alarmingly high estimates for bad data quality stemming from carelessness or other deceptive forms of behavior vary greatly between studies, methods, and recruitment methods. However, even a seemingly small number of careless responses can have serious consequences, such as failed replications (Oppenheimer, Meyvis, & Davidenko, 2009) or false-positives (Huang, Liu, & Bowling, 2015). Furthermore, careless responding may cause failed manipulations when instructions are not carefully read (Maniaci & Rogge, 2014), lower internal consistency of validated scales (Maniaci & Rogge, 2014), and problems in questionnaire development and item analysis (Johnson, 2005). Additionally, it can lead to problems in investigating questionnaire dimensionality (Kam & Meyer, 2015). However, estimates for careless behavior in crowdsourced samples remain unknown. All the aforementioned research examined academic participant pools or mixed types of online data (e.g., Maniaci & Rogge, 2014; Meade & Craig, 2012), or the studies only applied one measure to determine carelessness in a crowdsourced sample (Dogan, 2018; Hauser & Schwarz, 2016; Peer et al., 2017). Therefore, it was concluded there is a lack of a systematic analysis of careless behavior on crowdsourced platforms using various carelessness detection methods.

Recently, most of the attention of empirical research has been given to the discovery of carelessness (see Curran, 2016, for a review). The screening methods can be divided into two groups. The first group is the planned implementation of special items or scales to screen

carelessness. For example, Bogus Items (Meade & Craig, 2012), IRIs (Curran, 2016), and instructional manipulation checks (Oppenheimer et al., 2009). The second group of detection methods can be described as post hoc measures. These include the examination of response time, multivariate outliers, and (in-) consistency indices. These do not require special items, but an elaborate analysis after data collection. There are a variety of different methods available, but we will focus on those recommended in the recent literature (Curran, 2016).

Aim of the present study

Thus, the aim of the present study was to analyze the data quality of a crowdsourced online sample, based on various recommended methods for assessing careless behavior. This would address the limited variety of methods used in existing research about carelessness on crowdsourcing platforms.

Another open question revolves around the task-dependence of carelessness, and whether different methods embedded in different tasks capture different participants, or whether they stay careless for most of the study. As stated by Kan and Drummey (2018), it remains unclear in what way the duration or engagement level of a task impacts deceptive or careless behavior. Besides Likert-type scales, open questions (for example) are an extensively used method for capturing qualitative data. However, it is unknown how the quality of the answers given to such questions relates to carelessness.

To address these problems, the present study aims toward a better understanding of careless and inattentive behavior on crowdsourcing platforms (and the task-dependence of this phenomenon), by assessing the data quality with various detection methods. Moreover, we aim to provide pragmatic recommendations for ensuring survey data quality in research with crowdsourced samples. Based on these aims, we derived the following research questions:

Research Question 1: How prevalent is careless responding in samples from crowdsourcing platforms, based on various detection methods for carelessness?

Research Question 2: How are task-specific measures of carelessness (such as open-ended questions) related to planned detection methods and post hoc methods?

Research Question 3: Based on our findings, which methods are most applicable for

identifying carelessness in a crowdsourced sample?

Method

Data collection

The present study was conducted using a crowdsourced sample from FigureEight (CrowdFlower). Especially outside the U.S., FigureEight is a viable choice for crowdsourcing, as Amazon's MTurk has (for a long time) required requesters to have a US-address. FigureEight is accessible from Europe and other regions outside the USA, and provides access to millions of contributors (Van Pelt & Sorokin, 2012). The crowdsourcing platform is a well-established tool to gather participants for online-surveys, as shown by over 4600 hits on Google Scholar (21.03.2018, Keyword: CrowdFlower).
Data and analysis code used in this study is available at
`https://osf.io/9vjur/?view_only=9ed1707502684f89be168d358f5cd695`
(anonymized for peer review).

Procedure

After providing consent, participants were asked to recall a recent negative experience with an online store. In particular, participants were asked to respond to two questions 1) what exactly caused this experience to be negative and 2) how this affected their online shopping habits. We instructed participants to respond in free text with as much detail as possible, with complete sentences, and with at least 50 words. Next, 10 items of the positive and negative affect schedule (PANAS; Watson, Clark, & Tellegen, 1988), 23 items of the AttrakDiff2 (Hassenzahl, Burmester, & Koller, 2003), and 24 items measuring psychological need satisfaction adapted from Sheldon, Elliot, Kim, and Kasser (2001) were presented. This type of critical incident method is a common procedure in user experience research (e.g., Tuch, Schaik, & Hornbæk, 2016). After this first block of questions, participants were randomly allocated to be shown either a high trust or low trust mockup of a website. The website was manipulated according to the trust supporting elements identified by Seckler, Heinz, Forde, Tuch, and Opwis (2015). This setting was chosen to conduct a plausible experiment in user

experience research that was thematically related to the rest of the study. After this, participants were asked to complete 16 items of a Likert-type scale for trust in websites (Flavián, Guinalíu, & Gurrea, 2006). The goal of this section was to examine the effects of excluding data from careless participants on effect sizes and p-values in a group comparison. On the next page, participants rated the visual aesthetics of the website mock-up with 18 items (VisAWI, Moshagen and Thielsch (2010)). Following this section, the big five personality types were assessed with 44 items of the Big Five Inventory (BFI) (John & Srivastava, 1999). All post hoc detection methods of carelessness were investigated using this scale. On the last page of the survey, participants completed demographic information and a scale on self-reported careless responding (as in Maniaci & Rogge, 2014) and a self-reported single item (SRSI UseMe) (Meade & Craig, 2012). Finally, all participants were given a completion code.

Measures

All post hoc detection methods of carelessness were applied to the 44 items of the BFI in the last part of the questionnaire. We decided to focus on the BFI because it is multidimensional with a sufficient length to calculate various indices, and it is comparable with other studies in this field (Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012). The data of the other questionnaires used in this study were not subject to further analysis except for the trust scale by Flavián et al. (2006), which was used as a dependent variable in the experiment.

Planned detection methods

The wording of the self-reported responding tendencies scale, the Bogus Item, and the IRI incorporated in the study is presented in Table 1.

Self-reported responding tendencies. Following demographic questions, ten items based on Maniaci and Rogge (2014) were used to assess general tendencies in responding. Although excluding participants based on self-reported responding tendencies has been found to improve data quality significantly (Aust, Diedenhofen, Ullrich, & Musch, 2013), these items are also easily detected and prone to manipulation and dishonest answers. Three items were used to measure self-reported careless responding ($\alpha = .84$), two items to measure

self-reported patterned responding ($\alpha = .88$), three items to assess self-reported rushed responding ($\alpha = .83$), and two items assessing self-reported skipping of instructions ($\alpha = .68$). All items are presented in Table 1. Items were rated on a 7-point scale (1 = never, 4 = approximately half the time, 7 = all of the time), and responses were averaged ensuring high scores reflected more problematic responding.

Applying the cutoff used by Maniaci and Rogge (2014), answers of 4 or higher were flagged. Additionally, self-indicated data usage was assessed using the SRSI UseMe.

Attention checks. We employed two attention check items in the questionnaire following the Infrequency Approach (Huang, Curran, Keeney, Poposki, & DeShon, 2012), which entails including items to which all careful respondents should respond to in the same (or similar) fashion. One measure we applied was the Bogus Item similar to Meade and Craig (2012), which are items that are very unlikely for participants to agree with. The Bogus Item was located within the BFI (see Table 1). Participants who did not select "strongly disagree" or "slightly disagree" were thus flagged as inattentive. The other attention check item was an IRI similar to Meade and Craig (2012) and Curran (2016). According to Meade and Craig (2012) the IRI has several advantages over Bogus Items, as they are easier to create, have a singular correct answer, and therefore provide an obvious metric for scoring. Furthermore, they offer a clear interpretation and are not prone to humorous answers, which is a problem with the Bogus Item. The IRI (see Table 1) was placed within the items of the trust scale by Flavián et al. (2006). Participants who nevertheless answered this question were flagged.

Post hoc detection methods

Response Time. One simple post hoc measure to assess careless responding is to measure participant overall response time. The concept is that inattentive or careless respondents will be noticeable through unusually short or long completion times. Although this measure is easily applicable in any online survey, the issue of what constitutes an acceptable range of completion times must be decided individually for each question (Curran, 2016).

LongString Index. The LongString Index acts as an invariability measure, which assesses the number of same answers given in sequence. Careless participants who might select the

Table 1

The items of the self-reported responding tendencies scale (Maniaci & Rogge, 2014) and planned detection items included in the study. Self-report answer options ranged from 1 (never), over 4 (about half of the time) to 7 (all the time). The Bogus Item was included in the BFI where answers between 1 (disagree strongly) and 5 (agree strongly) were possible. The IRI was included in the trust scale that was used as the dependent variable of the experiment.

Measure	Item
Self-report	[How often do you...]
Careless responding	
	1. Read each question
	2. Pay attention to every question
	3. Take as much time as you need to answer the questions honestly
Patterned responding	
	4. Make patterns with the responses to a block of questions
	5. Use the the same answer for a block of questions one the same topic [rather than reading each question]
Rushed responding	
	6. Answer quickly without thinking
	7. Answer impulsively without thinking
	8. Rush through the survey
Skipping of instructions	
	9. Skim the instructions quickly
	10. Skip over parts of the instruction
SRSI UseMe	In your honest opinion, should we use your data in our analyses in this study? (Do not worry, this will not affect your payment, you will receive the payment code either way.)
Bogus Item	
	[I see myself as someone who ...] Did not read this statement
IRI	
	I read instructions carefully. To show that you are reading these instructions, please leave this question blank.

same answer for equal or greater than half the length of the total scale will be excluded from the sample (Curran, 2016; Huang et al., 2012). Curran (2016) recommended LongString analysis to identify some of the worst respondents that would otherwise be missed, although the measure can easily be deceived. The LongString Index in this study was calculated for the BFI following the procedure described in Meade and Craig (2012).

Odd-even consistency. To assess the Odd-even consistency (OEC), each individual's responses on each unidimensional subscale are split into responses to even and to uneven items (Curran, 2016). In the present case, this was implemented for each of the five dimensions of the BFI (Openness, Extraversion, Agreeableness, Conscientiousness, and Neuroticism). Reverse coded items must be recorded before calculating this measure. The responses to the even and uneven items are then averaged separately, ensuring each participant receives a score based on the even and the uneven items for each subscale of one larger scale. The individual correlation of these two vectors acts as a score of consistency. An important limitation is that this correlation is constrained by the number of subscales and the number of items in a scale. The OEC in this study was assessed for the BFI based on the procedure described by Meade and Craig (2012). Following Curran's (2016) recommendation, any correlation below 0 was flagged.

Resampled individual reliability. Curran (2016) proposed a more general conceptualization of the OEC measure – Resampled individual reliability (RIR). Here, the basic concept is that items that should measure the same construct should correlate positively within individuals. However, instead of limiting this idea to odd and even items, Curran (2016) suggests creating two halves of each subscale randomly without replacement. The individual correlation of these two vectors acts as a score of consistency. This process is then repeated several times (resampling). This is a new measure that was included in the present study and, to the best of our knowledge, has never been empirically examined. Following Curran's (2016) recommendation for the OEC, any correlation below 0 was flagged.

Person-total correlation. The measure of Person total correlation (PTC) describes the correlation of a participant,Äôs answers to each of the items of a scale, with the means of these items based on the whole sample (Curran, 2016). This measure relies on the assumption

that a large majority of the sample responded attentively, thus this measure may be problematic in situations where a large number of careless respondents are expected. Because this measure has currently not been empirically examined, no widely accepted cutoff value for this correlation exists. However, as recommended by Curran (2016), participants with a negative PTC were flagged.

Open answer quality. A priori criteria for the quality rating of open answers predominantly originates from the studies conducted by Holland and Christian (2009) and (Smyth, Dillman, Christian, & Mcbride, 2009). The following indicators for calculating an open answer quality index were taken into consideration: 1. Whether participants provided a thematically substantive response. 2. If a minimum of 50 words was provided (as instructed). 3. If participants provided answers in complete sentences (as instructed). 4. The number of subquestions answered (as instructed). 5. The number of subquestions further elaborated. A detailed description of how the open answer quality index was created is presented in the Appendix. The third author coded all experience reports. To ensure inter-rater reliability, the second author coded a random subset of 100 open-ended answers. Because two fixed raters rated a randomly selected subset, ICC3 was used (Koo & Li, 2016). Inter-rater agreement of each category was between moderate (Complete Sentences, $ICC3 = .80$), good (Substantive Response, $ICC3 = .78$; Number of Subquestions Elaborated, $ICC = .84$) and excellent (Number of Subquestions Answered, $ICC3 = .94$). Inter-rater agreement for the overall answer quality index was excellent $ICC3 = .96$, with a 95% confidence interval from .94 to .97 ($F(99,99) = 51$, $p < .001$).

Results

In this section, we first report on each group of carelessness detection methods separately, and then investigate how they relate to answer quality. Table 2 presents an overview of the number of participants flagged by each method.

Planned detection methods

Self-reported responding tendencies. Participants relatively frequently indicated that they engaged in problematic responding tendencies. Applying the cutoff used by Maniaci and

Table 2

Descriptive statistics for all detection methods used in the study. Self-report includes problematic responding tendencies as well as the SRSI UseMe item.

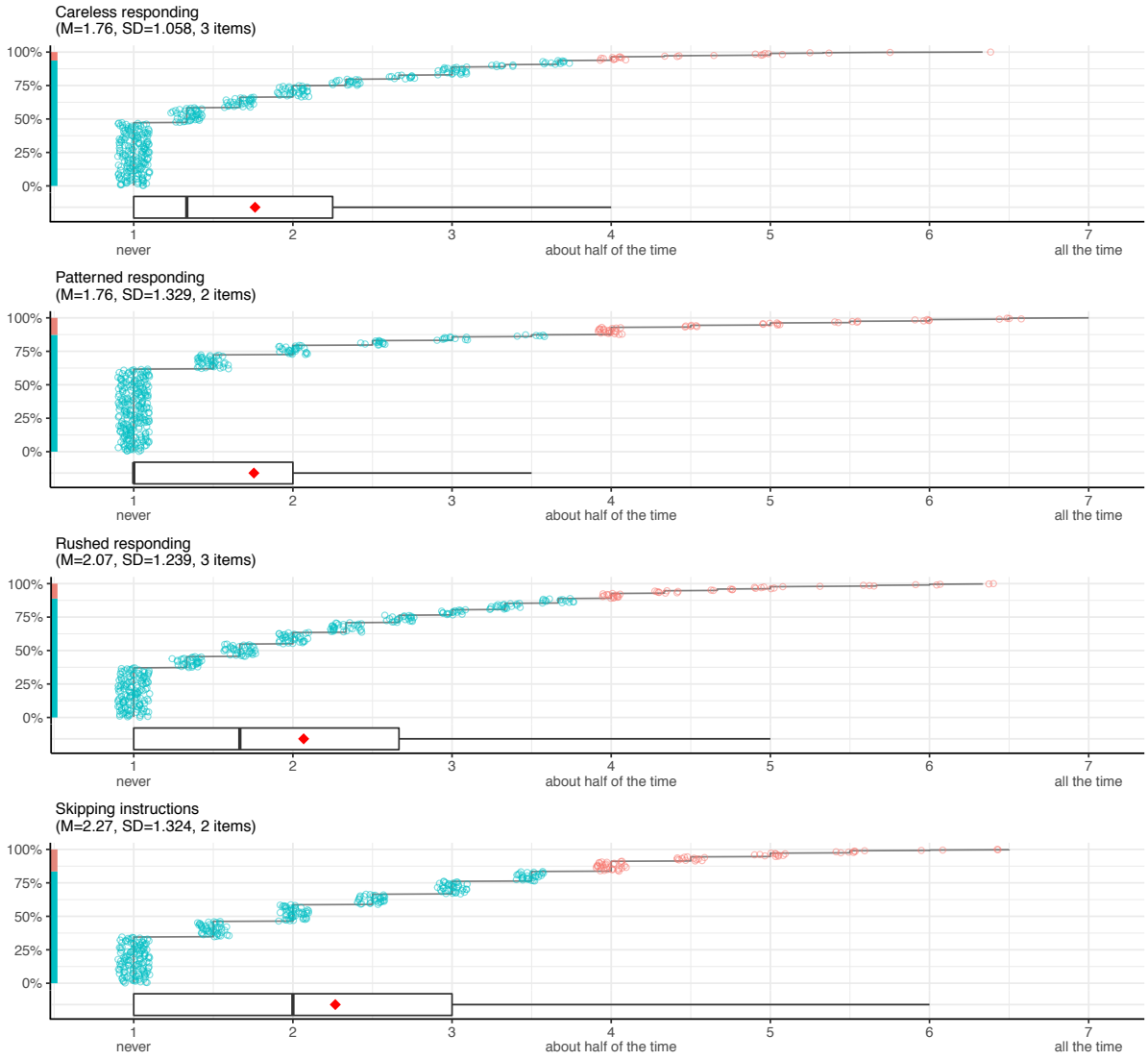
	Mean	SD	Min	Max	No. Flagged	%
Planned detection						
Self-report					106	26.90
Bogus Item					92	23.35
Instructed Response Item					96	24.37
Response time	16.71	9.22	3.93	61.15		
Post hoc detection						
LongString	6.63	9.15	0	44	25	6.35
Odd-even consistency	0.61	0.43	−1	1	63	15.99
Resampled individual reliability	0.56	0.39	−0.82	0.99	63	15.99
Person-total correlation	0.38	0.32	−0.47	0.88	74	18.78
Answer quality					100	25.38
Total (flagged by at least one method)					233	59.14

Note. Total $N = 394$

Rogge (2014), answers with 4 or higher were flagged. Thus, we flagged 25 careless respondents (6.6%), 50 pattern-respondents (12.7%), 44 rushed-respondents (11.2%), and 65 (16.5%) participants for skipping instructions. As depicted in Figure 1, skipping instructions was admitted most frequently (M = 2.27) followed by rushed responding (M = 2.07). However, there were fewer values of 4 and above for the rushed responding than for the patterned responding scale. Only 9 participants were flagged in every scale, 17 in 3 scales, 24 in 2 scales, and a majority of 49 participants were flagged in only 1 of the 4 self-reported scales. In total, the 4 scales flagged 99 (25.1%) participants as conspicuous.

The SRSI UseMe, indicating whether we should use the data provided by the participant or not, was negated by 22 participants (5.6%). Thus, these participants were also flagged as self-reported careless. It was then decided to aggregate these self-reported measures into one

Figure 1. Distributions of self-reported responding tendency scales. A random value was added to individual points to reduce overplotting.



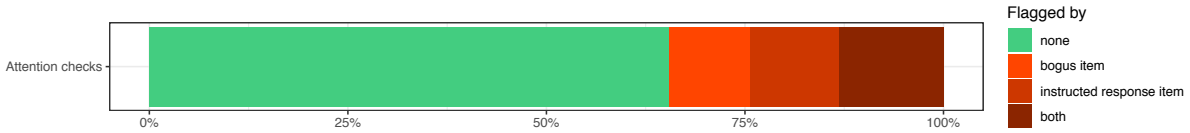
variable for self-reported carelessness.

Aggregating the self-reported measures of carelessness, patterned responding, rushed responding ($flagged >= 4$) and the SRSI UseMe, 106 participants (26.9%) were flagged as self-reported low-quality responses (see Table 2).

Attention checks. The IRI and the Bogus Item were missed by 96 participants (24.4%) and 92 participants (23.3%), respectively. Because there was no clear cutoff for the Bogus Item, we decided to code all answers with an agreement of 4 or higher to the item "*I see myself as someone who did not read this statement*" as failing to answer the Bogus Item correctly. Figure 2 demonstrates that the majority of respondents (258, 65.5%) answered both items

correctly, while 40 (10.2%) failed only at the Bogus Item, and 44 (11.2%) only at the IRI. However, a large number of participants who were flagged as inattentive missed both questions (52, 13.2%).

Figure 2. Number of participants flagged by one or both attention check items.



Post hoc detection methods

The boxplots and individual values of each post hoc detection method are presented in Figure 3. Where applicable, cutoffs are indicated by a vertical line and flagged participants are marked red ("fail") and inconspicuous participants are marked blue ("pass").

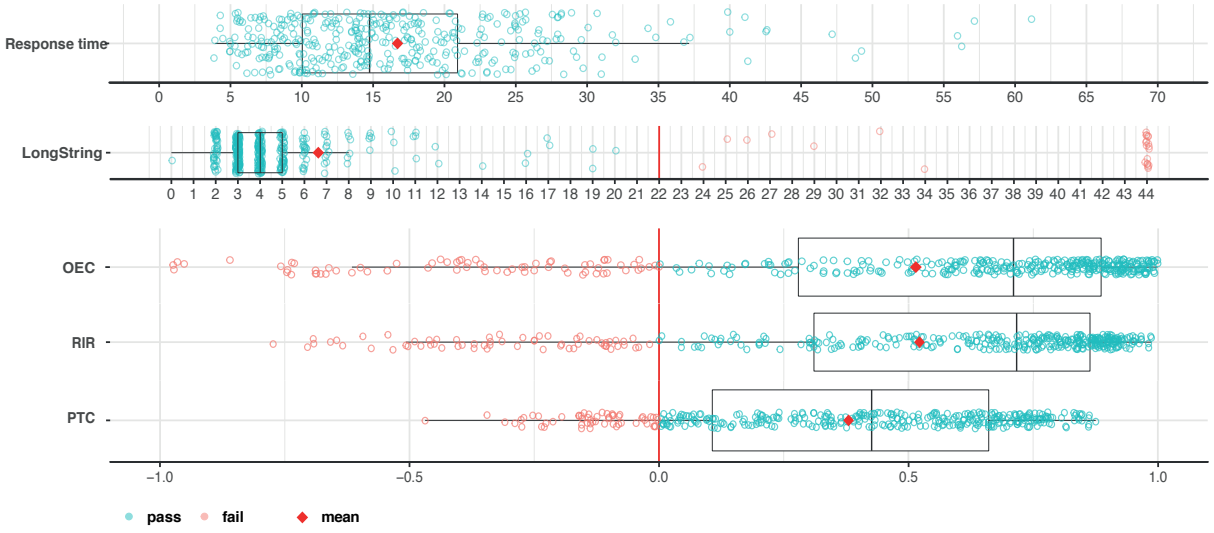
Response Time. Although Huang et al. (2012) recommended a general cutoff for too quick response times (2s an item), the distribution presented in Figure 3 did not show a cluster or conspicuous responses below a certain value. Therefore, no suspiciously fast respondents were flagged.

LongString Index. Results from the LongString analysis, with the recommended cutoff from Curran (2016) (>22), reveal that 22 (6.3%) of the participants were flagged by this method. The distribution depicted in Figure 3 displays that the vast majority of participants were significantly below this threshold, and 18 (4.6%) suspicious respondents with a LongString Index of 44 were identified with this method. These 18 participants provided the same answer for all 44 items of the BFI.

Odd-even consistency. The distribution in Figure 3 is left-skewed with a long tail, and only a few suspicious correlations are close to –1. Curran (2016) recommended removing all correlations below 0, which in this case would flag 63 (16.0%) of participants as responding too inconsistently.

Resampled individual reliability. As a more general approach to consistency than the OEC, RIR was calculated with 100 times randomly selected two halves of each subscale of the BFI. These two vectors were then correlated for each individual, giving a more general

Figure 3. Boxplots of carelessness detection methods. OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation. Response time in minutes for the entire survey.



(resampled) reliability. As with the OEC, the distribution is left-skewed with a long tail (see Figure 3). However, it has less extreme negative values, and slightly less respondents were identified as careless with this method (61, 15.5%).

Person-total correlation. Correlations of individual answers with the mean of answers from the whole sample exhibited a comparatively narrow distribution of values (see Figure 3). This method flagged 74 (18.8%) of participants as careless, indicated by a correlation of less than 0.

Open answer quality

Open answer quality was coded either 0 (= Insufficient), 1 (= High), or 2 (= Excellent). Of the full sample, 100 participants (25.4%) displayed insufficient answer quality in the open question. As we are mainly interested in whether participants failed or succeeded to provide sufficient open answer quality, the high (146, 37.1%) and excellent (148, 37.6%) open answer quality categories were combined for further analysis.

Relationship between open answer quality and carelessness detection methods. The 100 participants providing insufficient open answer quality will be referred to as the IAQ group in this section. Accordingly, the SAQ group represents the 294 participants with sufficient open answer quality. Results demonstrated that participants with IAQ significantly more often

failed the IRI ($\chi^2(1, N = 394) = 21.35$, $p < .001$) as well as the Bogus Item ($\chi^2(1, N = 394) = 24.665$, $p < .001$) than the SAQ group. Furthermore, 43% of all participants with IAQ were flagged by the self-reported carelessness measures, while only 21.4% were flagged in the SAQ group. The LongString cutoff flagged 15% of all IAQ participants and 3.4% of participants with SAQ. The 100 participants with low answer quality displayed higher LongString values (M = 9.55, SD = 12.42) than those with high quality (M = 5.63, SD = 7.49). A Wilcoxon rank-sum test yielded a significant difference at an alpha level of 5% (W = 17730, p < .01). Moreover, 24% of the IAQ group and 13.3% of the SAQ group failed the OEC cutoff. Similarly, 29% in group IAQ and 10.9% of group SAQ failed to display an RIR above the cutoff. For the PTC, 29% of group IAQ and 15.4% of group SAQ failed to display a positive correlation between their answer with the rest of the sample. Lastly, participants in the IAQ group (M = 882.58 seconds, SD = 563.66 seconds, n = 97) needed significantly less time to complete the survey than participants in the SAQ group (M = 1042.83, SD = 544.93, n = 289), in a Wilcoxon rank-sum test, W = 10958, p < .01.

Correlations between carelessness detection methods

Table 3 depicts how successfully the different methods correlate in their decision to classify participants either as suspicious or not suspicious. Answer quality achieved relatively low correlation with all behavioral and self-report measures of carelessness. The highest correlations of answer quality were observed with the Bogus Item (.26) and the IRI (.24). Interestingly, while the IRI and Bogus Item correlated with .41, the Bogus Item exhibited a higher correlation with the consistency measures PTC (.52), RIR (.51), LongString (.37), and OEC (.36). Self-reported data quality correlated substantially with RIR (.38), the Bogus Item (.34) and the IRI (.30). Unsurprisingly, the highest correlation was observed between OEC and RIR (.68), because RIR is a generalization of OEC. Overall, the correlation pattern demonstrates that among the attention check items the Bogus Item correlated more strongly with several other measures when compared to the IRI. The LongString Index exhibits similar correlations with all behavioral measures, except with IRI. The consistency measures correlate strongly with each other, apart from a relatively weak correlation between OEC and RIR (.25).

However, the relationship of answer quality with other measures is less clear. Based on these correlations, it is difficult to claim that one of the measures is redundant, as all the measures have relatively low overlap.

Table 3

Matthews correlation coefficient (MCC) of each measure pair (N = 394). A value near 1 suggests that the two methods have a high overlap in the classification of careless/not careless participants. IRI = Instruced Response Item, OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation.

	1.	2.	3.	4.	5.	6.	7.
1. Self-report	-						
2. Bogus Item	.34	-					
3. IRI	.30	.41	-				
4. LongString	.24	.37	.26	-			
5. OEC	.23	.36	.20	.40	-		
6. RIR	.38	.51	.22	.37	.68	-	
7. PTC	.31	.52	.27	.35	.25	.41	-
8. Answer quality	.21	.26	.24	.21	.13	.22	.15

Classification of respondents based on different methods.

Latent profile analysis. To identify different classes of carelessness, a Latent profile analysis (LPA) was conducted. Latent profile analysis is a flexible model-based approach to classification, with less restrictive assumptions than cluster analysis (Muthén, 2002). It aims to find the smallest number of profiles that can describe associations among a set of variables, and a formal set of objective criteria are applied to identify the optimal number of latent profiles in the data. For each participant, LPA provides a probability of membership, which is based on the degree of similarity with each prototypical latent profile. Following the approach by Meade and Craig (2012), we conducted an LPA on the non-self-report indicators of response quality (Open Answer quality, Response time, IRI, Bogus Item, LongString Index, OEC, RIR, and PTC) using the *mclust* package for R (Scrucca, Fop, Murphy, & Raftery,

2016). Self-report indicators were excluded, enabling a comparison of our results with Meade and Craig (2012), and because these indicators might be biased when participants are paid to participate. However, the self-report indicators were subsequently used to describe the different classes found in our data.

Open answer quality, IRI, and Bogus Item were binary variables (pass/fail). Missing data was present because for participants with a LongString Index of 44 (all items with the same answer) no OEC, RIR, and PTC measures could be computed (no variance in the answers). We therefore inputted missing values in these variables with +1 for consistency and reliability and –1 for PTC. Missing values in the response time variable were possible if participants did not respond to the questionnaire in one sitting. These missing values were estimated using an expectation maximization algorithm as implemented in mclust. Based on these variables, multiple models with one to nine classes were fitted. Bayesian information criterion (BIC) and integrated complete-data likelihood (ICL) criterion were used to judge the most appropriate number of classes. Both indicated that three classes were most appropriate (BIC: –7404.41, ICL: –7414.34). The class sizes were 181 (45.9%) for class 1, 129 (32.7%) for class 2, and 84 (21.3%) for class 3. The frequencies and variable means associated with each class are presented in Table 4.

As shown in Table 4, answers from class 1 were frequently judged as insufficient quality. Moreover, the attention check items were only missed by participants from this class. Further, class 1 participants more frequently self-reported bad quality than those in classes 2 and 3. Classes 1 and 2 responded significantly more quickly than class 3. Concerning OEC, class 3 provided more inconsistent answers than classes 1 and 2. Additionally, class 3 showed slightly stronger agreement to the self-reported responding tendencies than class 2. The defining hallmarks of class 1 were very large LongString Index values and very low PTC. This demonstrates that the consistency within participant answers was relatively high, while these answers were noticeably different from the total sample. Overall, it appears that a large part of class 1, which accounts for 45.9% of the sample, was responding in a careless way. However, class 1 cannot be described by one singular measure of carelessness. Instead, several forms captured by different methods should be included. In contrast, class 2 displayed the best values

Table 4

Descriptive statistics for each identified class of participants. IRI = Instruced Response Item, OEC = Odd-even consistency, RIR = Resampled individual reliability, PTC = Person-total correlation.

Variable	Class 1	Class 2	Class 3
Class size	181 (45.9%)	129 (32.7%)	84 (21.3%)
Percentages pass			
Answer quality (%)	44.75	100	100
Bogus Item (%)	49.17	100	100
IRI (%)	46.96	100	100
Self-report (%)	59.12	90.70	76.19
SRSI UseMe (%)	90.06	99.22	96.43
Means			
Response time (in Minutes)	14.58	16.94	22.03
OEC	.52	.86	.37
RIR	.44	.83	.43
PTC	.13	.59	.30
LongString	9.61	3.79	4.83
Means (Self-reported)			
Careless responding	2.16	1.36	1.52
Patterned responding	2.28	1.20	1.49
Rushed responding	2.43	1.68	1.88
Skipping instructions	2.53	1.99	2.13

for all examined measures. Class 3 was slightly more conspicuous in terms of self-reported scales, OEC, RIR, and PTC. This class appeared to answer slightly less consistently than class 2, but still managed to pass all attention checks and to provide sufficient open answer quality.

Prediction of class membership. It might not always be possible to incorporate all the above-mentioned carelessness detection methods in a study. Therefore, it was of interest to reduce the number of measures but still be able to identify participants of the careless class 1 accurately. Conditional inference trees, as implemented in the *party* package for *R* (Hothorn, Hornik, & Zeileis, 2006), were used to identify the most predictive measures for class membership for each participant. Conditional inference trees use a recursive algorithm to make an unbiased selection among covariates, and offer several advantages over traditional regression models and random forests (Hothorn et al., 2006; Strobl, Malley, & Tutz, 2009), such as non-linear relationships and less overfitting. Nine variables were used to predict class membership (SRSI UseMe, Bogus Item, IRI, Response time, LongString, OEC, RIR, PTC, and Open answer quality). The SRSI UseMe variable, Bogus Item, IRI, and the answer quality were included as binary variables (Pass/Fail), whereas the remaining variables were used in their raw form. Results of the analysis depicted in Figure 4 demonstrate that answer quality, IRI, and Bogus Item are well suited to separate the careless class 1 from classes 2 and 3. Furthermore, taking post hoc detection methods such as LongString analysis, OEC, PTC, and Response time into account, the tree successfully separates classes 2 and 3. Table 5 demonstrates that the prediction based on this model is very accurate (Accuracy = .987, 95% CI[.971, .996]) in terms of identifying the correct class membership. Only 5 participants out of 394 were assigned to the wrong class based on this model.

Effects of carelessness on experimental manipulation

The goal of the experiment included in the study was to examine how effect sizes and p-values changed when careless participants were excluded from the analysis. Results of a Welch's t-test with the full sample demonstrated that there was a significant difference in perceived trustworthiness of the online shop, $t(381.83) = 5.64$, $p = 3.344e-08$, $d = 0.567$. Participants who saw the low-trust website mock-up rated the company as less trustworthy (M = 4.36, SD

Figure 4. Conditional inference tree for all carelessness detection methods. For each inner node, the Bonferroni-adjusted p-values are presented, the fraction of participants in each class (1, 2, or 3) is displayed for every terminal node.

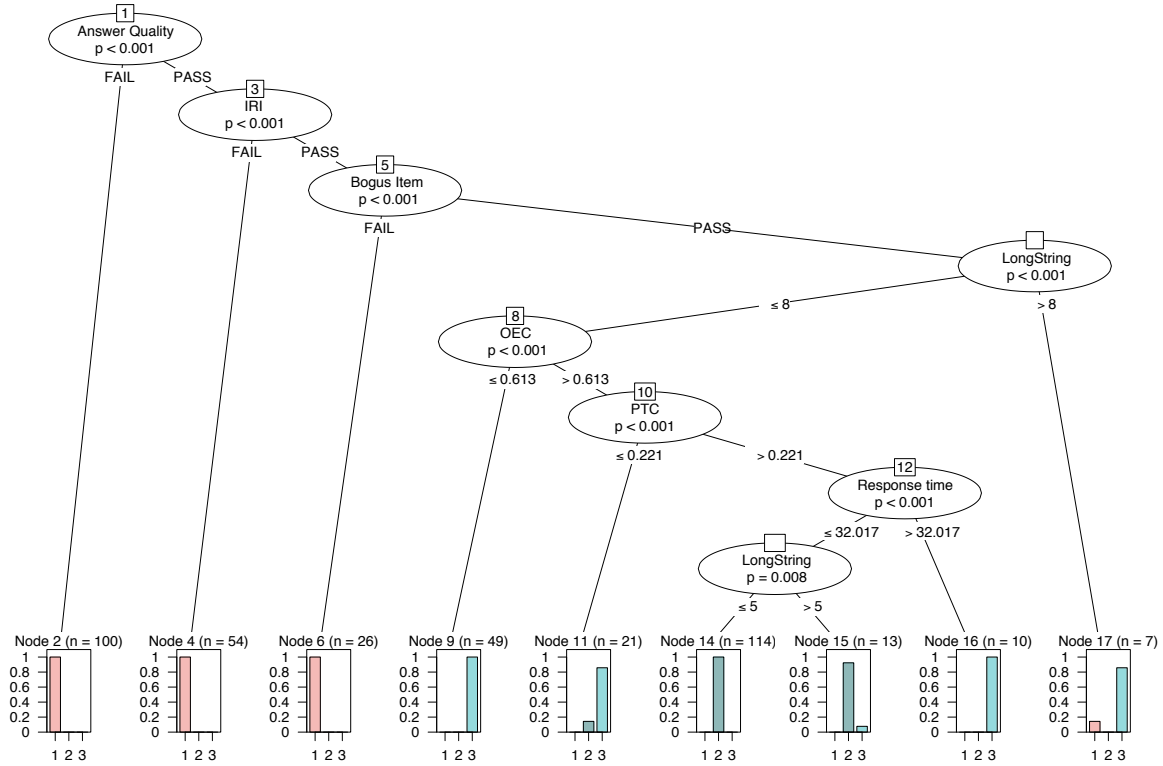


Table 5

Performance of the conditional inference tree model in predicting class membership.

	Class 1	Class 2	Class 3
Predicted			
Class 1	180	0	0
Class 2	0	126	1
Class 3	1	3	83

Note. $N = 394$

= 1.21) than participants in the high-trust condition (M = 4.99, SD = 1). When participants from class 1 (n = 181) were removed, participants in the low-trust condition rated the company slightly less trustworthy (M = 4.34, SD = 1.19), and participants in the high-trust condition rated the website somewhat more trustworthy than the full sample (M = 5.12, SD =

0.95). The standard deviations decreased slightly in both groups, which indicates that some of the noise that could stem from careless participants was reduced. Although the differences between the two conditions was significant in both cases, removing participants from the careless class 1 led to a smaller p-value and increased the effect size, $t(194.32) = 5.83$, $p = 2.277e-08$, $d = 0.803$.

Discussion

Analysis of careless behavior in a crowdsourced sample

Previous work studied carelessness or other deceptive forms of behavior in online samples either with only a few methods (J. J. Chandler & Paolacci, 2017; Hauser & Schwarz, 2016; Kan & Drummey, 2018; Peer et al., 2017), or they assessed carelessness in student samples or mixed online samples (Maniaci & Rogge, 2014; Meade & Craig, 2012). We build on this work with a systematic analysis of carelessness in a crowdsourced sample and examine the new methods: RIR and PTC (Curran, 2016). We applied six measures and corresponding cutoffs, based on recommendations from Meade and Craig (2012), Maniaci and Rogge (2014), and Curran (2016), to identify multiple forms of carelessness in a crowdsourced sample from FigureEight.

Observing the planned detection methods, which require special items or scales, 26.9% of all participants indicated in self-reports to provide careless, patterned, or rushed responses, 24.4% of all participants failed to answer the IRI correctly, and 23.4% missed the Bogus Item (see Table 2). Weak to moderate correlations between aggregated self-reported carelessness and other detection methods only partially indicate convergent validity for self-report measures (see Table 3). Correlations between attention check items and other detection methods were also weak to moderate, except for the Bogus Item that correlated relatively strongly with RIR and PTC. The 24.4% of participants in our FigureEight sample who failed the IRI surpass the 14% found in the study by Maniaci and Rogge (2014), which examined a sample including MTurk workers, participants from online forums, and psychology students. This indicates that inattentive behavior may be more frequent in samples from crowdsourcing platforms. Taken together, approximately 25% of the sample was flagged as inattentive, based solely on one of

the attention check items. It can be expected that the overall number of participants flagged with these items increases with the length of the survey, as one attention check item is recommended for every 50–100 items (Meade & Craig, 2012). In a considerably longer study, applying 4 attention check items, Peer et al. (2017) found 73% of all participants fail at least one attention check item. Post hoc detection methods revealed 6.3% of all participants were flagged by the LongString analysis, which corresponds with findings from Maniaci and Rogge (2014), where 6% were flagged in a mixed sample. However, the OEC and RIR revealed 16% and 15.5%, respectively, as responding too inconsistently. This is more than twice as much as Maniaci and Rogge (2014) identified with the OEC method. Lastly, the PTC flagged 18.8% of all participants as being careless. Hence, the post hoc detection methods of the present study further suggest that careless or inattentive behavior may be more pronounced in a fully crowdsourced sample.

How prevalent is careless responding in samples from crowdsourcing platforms, based on various detection methods for carelessness? (RQ1)

Almost 60% of all participants were flagged by at least one of the methods examined in this study (see Table 2). However, the univariate examination of single measures, and a subsequent cumulative exclusion of participants, might be problematic for various reasons. Firstly, with this strategy, participants are excluded based on methods that do not have a set cutoff or an objective wrong answer, and the researcher has to decide whether one or multiple flags per participant would lead to an exclusion from the sample. Secondly, simply combining the different measures altogether might be too restrictive and lead to many false-positives. For instance, the PTC cutoff might not be meaningful in situations where a lot of carelessness can be expected. Therefore, and in line with Maniaci and Rogge (2014) and Meade and Craig (2012), we base our prevalence estimate of carelessness on the results of the LPA, which takes multiple raw values of various non-self-report methods into account to identify different classes of participants.

The LPA identified three classes in total. Class 1 (the careless participant class), contained 45.9% of all participants. Although this class cannot be described by one measure, and

therefore comprises multiple forms of inattention and carelessness, its characteristics can be summarized as follows: Failing in providing sufficient open answer quality and failing attention checks was an exclusive characteristic of this class. This class also self-indicated bad data quality considerably more often than the other two classes. Participants of this class answered more quickly, showed very large LongString values, and a very low PTC, indicating excessive consistency within, yet low congruence with the total sample. While the OEC measure also revealed a relatively high inconsistency in the answers of this class, it is important to note that class 3, usually inconspicuous concerning other detection methods of carelessness, provided even more inconsistent answers. This finding suggests using caution with measures of consistency as a means of data cleaning, because they might bear potential for a high false-positive rate. The LPA from the present study revealed a considerably larger group of careless participants (45.9%) in a crowdsourced sample compared to similar analyses conducted with mixed online samples or student pools in the studies in Maniaci and Rogge (2014) and Meade and Craig (2012). These studies identified approximately 2.2% to 11% as being careless. Concerns surrounding the representativeness of a sample after excluding such a large percentage of participants, and from an economic perspective, might suggest not using such a sample.

How are task-specific measures related to planned detection methods and post hoc methods? (RQ2)

Out of 394 participants, 100 (25.4%) provided insufficient open answer quality. Significantly fewer participants of this group passed attention checks; they more often self-reported bad data quality and they exhibited significantly higher LongString Index values. Furthermore, participants who failed in providing sufficient answer quality completed the survey in significantly less time, they more often failed to meet the OEC cutoff and the RIR, and they more often failed to meet the PTC cutoff. Hence, these results indicate some convergent validity for open answer quality as a measurement for carelessness. However, correlations between this measure and other planned detection or post hoc methods were rather weak (see Table 3). This might indicate that carelessness depends (to a large extent) on the given task.

This result coincides with findings from Maniaci and Rogge (2014), indicating that inattention or carelessness during specific tasks (such as watching a video or marking pronouns in a text), mostly has a low correlation with other detection methods of carelessness. Therefore, completing standardized Likert-scale questionnaires, answering open questions, watching videos, and participating in concentration tasks in online studies appears to evoke different forms of inattention, which tend to concern different participants.

Which methods are most applicable to identify carelessness in a crowdsourced sample can be made, based on our findings? (RQ3)

In general, we strongly encourage other researchers to analyze the data quality of crowdsourced surveys. As in Maniaci and Rogge (2014) and Meade and Craig (2012), we refer to the LPA as our reference for careless behavior in our sample. Based on our findings, we recommend a set of measures that are easy to apply, easy to interpret, and at the same time cover the majority of the inattentive class 1.

1. Further, we recommend including an SRSI UseMe item to assess whether participants indicate that their data should be used for the study. Although this item was not an important predictor of class membership, it acts as a form of revoked consent. Thus, it serves a purpose beyond detecting bad data quality. However, from a practical perspective, we cannot currently recommend other self-report measures. This is because including 10 or more additional items in a survey with the sole purpose of detecting self-reported bad data quality may not be an efficient approach for all online surveys.

2. Attention checks such as an IRI should be included, because these detection methods are easy to create and offer a clear interpretation. We further advise to include a Bogus Item, as the combination of a task-specific measure, the IRI, and the Bogus Item was successful in classifying 180 of 181 participants correctly in class 1. However, the wording of the Bogus Item should be chosen carefully, because Bogus Items can cause interpretative problems (Meade & Craig, 2012).

3. The inclusion of the LongString Index as a post hoc measure is recommended, because this measure is applicable to all type of scales (given sufficient length), and provides an

overview of repetitive answer patterns. Moreover, a high LongString Index was a typical characteristic of the inattentive class 1 (see Table 4). However, in most cases, a task-specific measure and a combination of attention checks appears sufficient, as the LongString Index was not a significant predictor for class 1 in the conditional inference tree. Based on the results of the conditional inference tree and the LPA, we cannot recommend the other post hoc detection methods. The minor response time differences between classes 1 and 2 did not offer a clear and readily applicable cutoff value for an inattentive class (see Table 4). Furthermore, response time was not identified as a significant predictor of class 1. Concerning OEC, the LPA identified a class (class 3, see Table 4), with slightly lower values than the careless class 1. However, apart from this measure, and a considerably longer average response time, this class was inconspicuous. Thus, flagging participants based on this measure might lead to a high false-positive rate. The RIR of class 1 was comparable to class 3. However, although class 1 showed a lower PTC (see Table 4), this measure was not predictive for class 1 in the model. In comparison to a LongString analysis, the interpretation of this measure is more dependent on the properties of a sample. In a sample with poor data quality, this measure is heavily biased, because it correlates individual responses with averaged responses, including all careless participants (Curran, 2016).

4. Results suggest that carelessness is dependent on the given tasks to participants. The correlation table (see Table 3) demonstrates that the open answer quality achieved relatively low correspondence with other detection methods for carelessness. Meanwhile, low open answer quality is exclusively (and thus clearly) associated with the inattentive class 1. Therefore, we encourage researchers to apply carelessness detection methods according to the given tasks. While planned and post hoc detection methods might generally identify carelessness in Likert-type questionnaires, they might not prevent bad data quality in other types of online-survey tasks.

Taken together, we recommend the following set of carelessness detection methods: an SRSI UseMe item, one or multiple Instructed Response or Bogus Items, a LongString analysis, and a task-specific measure (in our case: assessing open answer quality). These measures either

represented important predictors for the inattentive class 1, or they provided pragmatic merit for analyzing the data quality of a crowdsourced sample. All these methods are relatively straightforward to apply, as they do not need to consider scale dimensions and inverse items. Furthermore, they were clearly associated with the inattentive class 1 in the LPA, and the prediction for class 1 (based on these detection methods) was very accurate: 180 out of 181 were correctly identified, while none of the participants from classes 2 and 3 were falsely flagged by this combination of methods. As demonstrated by the experiment included in this study, removing careless participants increased effect sizes from d = 0.567 to d = 0.803. Although the difference was very robust in the sample including careless respondents, research has also shown that carelessness can not only reduce effects but also disperse known effects (e.g., DeSimone & Harms, 2018; Maniaci & Rogge, 2014). Hence, carelessness may reduce statistical power and increase noise in the data, thus undermining the validity of online experiments. Therefore, it is vital that researchers develop a data cleaning strategy whenever online samples are recruited, and cleaning process must be reported in detail.

Limitations and future research

Some limitations must be considered concerning the results and recommendations presented in this paper:

First, the present study was conducted on the FigureEight platform, and this might not readily translate to other platforms or recruitment methods. For instance, Amazon's MTurk offers different methods of community-management and rating possibilities for workers, which may cause workers to be more attentive when taking part in a survey. Future research, therefore, should systematically assess data quality differences between various platforms, applying multiple carelessness detection methods.

Second, the present study assessed the detection of careless participants, which resulted in the exclusion of approximately half of all participants. Excluding this number of participants could have severe methodological and financial implications. Hence, future research should also focus on preventing carelessness, which is presently not well understood. Warnings about monitoring data quality that have been used by Clifford and Jerit (2015) or Meade and Craig

(2012) can be effective, but might lead to other biases, such as socially desirable behavior. D. Chandler and Kapelner (2013) have found positive effects of meaning by explaining the purpose of a task on data quality in crowdsourcing tasks. Furthermore, Ward and Pond (2015) found that promising participants the results of the study was effective in increasing data quality. More effort is needed to systematically analyze these measures for preventing carelessness in crowdsourced samples.

Third, the present study included an open-ended question to assess task-dependency of careless behavior. Although findings from Maniaci and Rogge (2014) suggested a similar approach (by applying other forms of different tasks in their survey), future research should aim for a systematic review of a wider variety of different tasks in online surveys. This will facilitate further analysis of the task-dependency of careless behavior.

Finally, as all post hoc detection methods are approximate and uncertain, bad data quality can not clearly and reliably be identified in every case. Our recommendations are based on the prediction of class 1, which was identified using LPA. Only planned detection methods were found to be predictive for this class. However, there are situations where it might not be possible to include attention check items or task-dependent measures of quality, such as voluntary surveys of highly specific populations. Hence, further research is needed to ensure data quality in such situations.

Conclusion

The aim of this study was to provide an estimate of the frequency of carelessness in samples from crowdsourcing platforms, based on different identification methods. Our results reveal that approximately half of all crowdsourced participants display careless behavior. Furthermore, carelessness and inattention appear highly task-dependent, as correlations between open answer quality and other measures are rather low. Finally, based on a predictive model and interpretative problems of several detection methods, we recommend assessing data quality of crowdsourced samples by applying the following: an SRSI UseMe item, attention checks such as the IRI and Bogus Item, the LongString Index, and a task-specific measure. A combination of these methods was able to identify 180 out of 181 inattentive

participants, and the subsequent exclusion of this subsample resulted in an increased effect size and smaller p-value in the experiment.

Acknowledgments

This work has been approved by the Institutional Review Board of the Faculty of Psychology, University of Basel under the number D-006-17. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. doi: 10.3758/s13428-012-0265-2

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. doi: 10.1016/j.chb.2013.05.009

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133. doi: 10.1016/j.jebo.2013.03.003

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. doi: 10.3758/s13428-013-0365-7

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, *26*(7), 1131–1139. doi: 10.1177/0956797615585115

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*(1), 53–81. doi: 10.1146/annurev-clinpsy-021815-093623

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500–508. doi: 10.1177/1948550617698203

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon mechanical turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361. doi: 10.1007/s10869-016-9458-5

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, *79*(3), 790–802. doi: 10.1093/poq/nfv027

Comley, P. (2015). Online market research. In M. v. Hamersveld & C. d. Bont (Eds.), *Market*

Research Handbook (pp. 401–419). Chichester, England: John Wiley & Sons Ltd. doi: 10.1002/9781119208044.ch21

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. doi: 10.1016/j.jesp.2015.07.006

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, *33*(5), 559–577. doi: 10.1007/s10869-017-9514-9

de Winter, J., Kyriakidis, M., Dodou, D., & Happee, R. (2015). Using crowdflower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturing*, *3*, 2518–2525. doi: 10.1016/j.promfg.2015.07.514

Dogan, V. (2018). A novel method for detecting careless respondents in survey data: floodlight detection of careless respondents. *Journal of Marketing Analytics*, *6*(3), 95–104. doi: 10.1057/s41270-018-0035-9

Douglas, K. M., & McGarty, C. (2001). Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology*, *40*(3), 399-416. doi: 10.1348/014466601164894

Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management*, *43*(1), 1–14. doi: 10.1016/j.im.2005.01.002

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in mturk and other online samples. *Industrial and Organizational Psychology*, *8*(2), 196–202. doi: 10.1017/iop.2015.25

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1631–1640). New York, NY, USA: ACM. doi: 10.1145/2702123.2702443

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*(1), 877-902. doi: 10.1146/annurev-psych-010814-015321

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung

wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Wiesbaden: Vieweg+Teubner Verlag. doi: 10.1007/978-3-322-80058-9_19

Hauser, D. J., & Schwarz, N. (2016, Mar 01). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. doi: 10.3758/s13428-015-0578-z

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review*, *27*(2), 196-212. doi: 10.1177/0894439308327481

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. doi: 10.1198/106186006X133933

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. doi: 10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828. doi: 10.1037/a0038510

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research* (Vol. 2, pp. 102–138). New York, NY, US: Guilford Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. doi: 10.1016/j.jrp.2004.09.009

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. doi: 10.1177/1094428115571894

Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? an examination of worker misrepresentation and downstream consequences in amazon's mechanical turk

workforce. *Computers in Human Behavior*, *83*, 243–253. doi: 10.1016/j.chb.2018.02.005

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. doi: 10.1016/j.jcm.2016.02.012

Lee, H. (2006). Privacy, publicity, and accountability of self-presentation in an on-line discussion group. *Sociological Inquiry*, *76*(1), 1–22. doi: 10.1111/j.1475-682X.2006.00142.x

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. doi: 10.1016/j.jrp.2013.09.008

McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior*, *84*, 295 - 303. doi: 10.1016/j.chb.2018.03.007

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. doi: 10.1037/a0028085

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709. doi: 10.1016/j.ijhcs.2010.05.006

Muthén, B. O. (2002). Beyond sem: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117. doi: 10.2333/bhmk.29.81

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. doi: 10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. doi: 10.1016/j.jesp.2009.03.009

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi:

10.1177/0963721414531598

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. doi: 10.1016/j.jesp.2017.01.006

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205–233.

Seckler, M., Heinz, S., Forde, S., Tuch, A. N., & Opwis, K. (2015). Trust and distrust on the web: User experiences and website characteristics. *Computers in Human Behavior*, *45*, 39—50. doi: 10.1016/j.chb.2014.11.064

Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, *80*(2), 325. doi: 10.1037/0022-3514.80.2.325

Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology*, *57*(1), 529-555. doi: 10.1146/annurev.psych.57.102904.190048

Smyth, J. D., Dillman, D. A., Christian, L. M., & Mcbride, M. (2009). Open-ended questions in web surveyscan increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*(2), 325–337. doi: 10.1093/poq/nfp029

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736 - 748. doi: 10.1016/j.tics.2017.06.007

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. doi: 10.1037/a0016973

Tuch, A. N., Schaik, P. V., & Hornbæk, K. (2016). Leisure and work, good and bad: The role of activity domain and valence in modeling user experience. *ACM Transactions on Computer-Human Interaction*, *23*(6), 35:1–35:32. doi: 10.1145/2994147

Van Pelt, C., & Sorokin, A. (2012). Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of*

Data (pp. 765–766). New York, NY, USA: ACM. doi: 10.1145/2213836.2213951

Ward, M., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417 – 430. doi: https://doi.org/10.1016/j.chb.2017.06.032

Ward, M., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554–568. doi: 0.1016/j.chb.2015.01.070

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063. doi: 10.1037/0022-3514.54.6.1063

Appendix

Calculating the open answer quality Index

A-priori criteria for the rating of open-ended questions were defined according to the measures used in studies from Holland and Christian (2009); Smyth et al. (2009). The following indicators for calculating an open answer quality index were taken into consideration:

Substantive response. This indicator refers to whether the participant answer thematically corresponds to the open question subject matter. The open answer has been coded with 0 if it merely consisted of meaningless sequences of letters, clearly copy-pasted phrases, or thematically unfit answers which typically emerged from not carefully reading the instructions (such as describing a negative experience in a non-virtual store instead of an online shop). If the open answer corresponded to the subject matter, regardless whether the participant addressed all subquestions, the indicator has been coded with 1.

Number of words. Because it is possible to provide a thematically substantial answer while providing little or zero actual content (such as merely writing one short sentence about the experience), the number of words has been assessed for each open answer. Given the topic of the open question and the two subsequent subquestions, a minimum of 50 words ($\pm$3) was defined as the requirement to answer the questions. Thus, participants were explicitly asked to

provide an answer containing at least 50 words. This number is regarded as being a minimum effort to achieve a thematically substantial answer that additionally addresses at least one subquestion. Wordcounts corresponding to this number (or higher) were coded with 1, smaller wordcounts with 0.

Complete sentences. Participants were explicitly asked to provide answers with full sentences. Open answers that mainly or exclusively consisted of unfinished sentences (or separate words) were coded with 0 in regard to complete sentences. To receive a coding of 1, the majority of all sentences in the open answer needed to be complete and separated with commas or periods.

Number of subquestions answered. If none of the specific subquestions were addressed, the answer was coded with 0 in regard to number of subquestions. This was also the case if the given answer met the requirements for a thematically substantial answer, but failed to answer at least one of the specific subquestions. Accordingly, the answer received a coding of 1 or 2 if one or both subquestions were addressed in the open answer, respectively.

Number of subquestions elaborated. An answer to a subquestion was considered to be elaborate if the according part of the open answer contained at least three complete sentences. If none of the subquestions were elaborated, the answer was coded with 0 in regard to themes elaborated. Accordingly, the answer received a coding of 1 or 2 if one or both subquestions were elaborated in the open answer, respectively.

Calculation of the open answer quality Index. *Substantive response* and *Number of words* were seen as essential for providing a valuable open answer. Thus, if one or both of these variables were coded with 0, the open answer quality Index was also automatically coded with 0. The other variables, namely *complete sentences*, *number of subquestions answered* and *number of subquestions elaborated*, were seen as being important (but not an absolute necessity) on their own in order to provide a good open answer quality. Thus, for answers that met the minimum requirements, the codings from *complete sentences*, *number of themes*, and *themes elaborated* were counted together. If the sum reached 3 or higher, the overall open answer quality was considered to be adequate, and thus coded with 1.

TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web

Florian Brühlmann, Serge Petralito, Denise C. Rieser, Lena F. Aeschbach, Klaus Opwis

Center for Cognitive Psychology and Methodology, Department of Psychology, University of Basel

Abstract

Trust is an essential factor in many social interactions involving uncertainty. In the context of online services and websites, anonymity and lack of control make trust a vital element for successful e-commerce. Despite trust receiving ongoing attention, there is a need for validated questionnaires that can be readily applied in different contexts and with various products. We therefore report the development and validation of a semantic differential measuring users' trust on three dimensions. Compared to Likert-type scales, semantic differentials have advantages when it comes to measuring multidimensional constructs in various contexts. The TrustDiff measures users' perceptions of Benevolence, Integrity, and Competence of an online vendor with ten items. The scale was investigated in three independent studies with over 1000 participants and shows good structural validity and high reliability, and correlates as expected with related scales. As a test of criterion validity, the TrustDiff showed significant differences on all subscales in a study involving a manipulated website.

Keywords: Trust, Semantic differential, Scale development, User experience, E-commerce

1. Introduction

Trust is an essential factor when acting under uncertain conditions and with the risk of negative consequences (Casaló et al., 2007). There are multiple definitions of trust in the literature, emanating from various academic fields (e.g., Driscoll, 1978; Moorman et al., 1993; Rotter, 1967). This renders a precise operationalization for measuring trust particularly challenging. Most definitions have two key components of trustworthiness in common: a willingness to be vulnerable, and a perception of the intentions of the other party (Lewicki

and Brinseld, 2012). The concept of trust in the web has important differences from trust in offline contexts. Online trust is usually complicated by trust in the internet itself, and the organization behind the technology. In addition, trust is characterized by a lack of face to face interaction, an asymmetry in the information available to each party, and concerns about privacy (van der Werff et al., 2018). The question of whether trust in a web context refers to the organization behind a website, to individuals (who, for example, will select or deliver your order), or to the internet technology itself (such as online payments) is still open for debate (van der Werff et al., 2018).

However, trust in a web context is usually built on characteristics from e-commerce (Wang and Emurian, 2005). Accordingly, several questionnaires have been developed to measure trust (e.g., Bhattacherjee, 2002; Cho, 2006; Flavián et al., 2006; Gefen, 2002; McKnight et al., 2002). One of the main issues of trust research in web or e-commerce contexts is the lack of a common, validated, reliable, and versatile measure (Kim and Peterson, 2017). We further identify several limitations of the abovementioned scales with regard to applicability in research and practice. First, most questionnaires incorporate Likert-type scales with domain-specific statements. For instance, the items developed by McKnight et al. (2002) are tailored to a specific website under examination (such as "LegalAdvice.com is competent and effective in providing legal advice"). In order to apply the scales in a different context, it may be necessary to rephrase its items. However, rephrasing the statements used in these questionnaires could result in a loss of reliability and validity. Second, translating Likert-type statements into other languages can be a difficult and time-consuming process, which may further affect validity.

In the present work, therefore, we describe the development and validation of TrustDiff, a semantic differential for measuring trust in the web. This new measure displays several advantages over traditional Likert-type scales when measuring complex and multidimensional constructs (Verhagen et al., 2015). The results of three validation studies (total sample size $N = 1165$) indicate that the TrustDiff has excellent psychometric properties, measuring Benevolence, Integrity, and Competence with high reliability. Furthermore, we demonstrate how these three subscales relate to an existing Likert-type trust scale and the concepts of

visual aesthetics and usability. Finally, the TrustDiff was found to be sensitive to the manipulation of trust-related features in an experiment with a mock website. Taken together, the TrustDiff represents a promising tool for assessing trust in various domains of research and practice.

1.1. Characteristics and dimensions of trust

There are four characteristics of trust that are generally observed and agreed upon in the context of trust in e-commerce (Wang and Emurian, 2005). First, there must be two specific parties in a trusting relationship: a trusting party (truster, such as an online customer) and a party to be trusted (trustee, such as an online merchant). Second, trust involves vulnerability, uncertainty and risk for the truster, while anonymity and unpredictability are associated with the trustee. Third, trust leads to actions that mostly comprise risk-taking behaviors, such as providing personal and financial information. Finally, trust is subjective, and the level of trust considered sufficient for online transactions is different for everyone. Moreover, people vary in their attitudes toward machines and technology (Wang and Emurian, 2005). Trust in e-commerce involves interpersonal trust, trust in the organization representing a website, and trust in the underlying technologies (van der Werff et al., 2018). In the Web Trust Model developed by McKnight et al. (2002), trusting beliefs are at the core of what we consider the different dimensions of user trust.

Although there are multiple types of trusting beliefs found in the literature, three dimensions are generally accepted (Bhattacherjee, 2002; Chen and Dhillon, 2003; Flavián et al., 2006; Gefen, 2002; Mayer et al., 1995; McKnight et al., 2002): *Benevolence*, *Integrity* and *Competence*. Benevolence is related to the users belief that the other party is interested in their welfare, is motivated by a search for a mutually beneficial relationship, and has no intention of engaging in opportunistic behavior. Integrity, sometimes referred to as honesty Flavián et al. (2006), is the belief that the other party is sincere and fulfills its promises. Finally, Competence implies that the other party has the resources and capabilities needed for the successful completion of the transaction, and for the continuance of the relationship (Casaló et al., 2007).

1.2. Existing questionnaires

Various works have been directly or indirectly concerned with measuring trust (Bart et al., 2005; Cho, 2006; Corbitt et al., 2003; Lee and Turban, 2001; Jarvenpaa et al., 1999; McKnight et al., 2002; Pavlou and Gefen, 2004). However, from the practical and research perspectives, there remains a need for a validated, brief, and easy-to-translate scale that measures trust and incorporates the three dimensions of Benevolence, Integrity, and Competence (Kim and Peterson, 2017). The following problems with preexisting scales have been identified: first, not all existing scales inquire about trust directly; they often ask about adjacent constructs such as Benevolence in Cho (2006), which in McKnight et al. (2002) is merely a part of the model for trust. Second, existing measurement methods have been created to answer specific questions in certain contexts. An example of this is Lu et al. (2012), who developed Likert-type questions for customer-to-customer (C2C) platforms, such as "Do you agree that this C2C platform solves a security problem or stops fraudulent behavior?". Third, in their meta-analysis, Kim and Peterson (2017) described preexisting measurements as "ambiguous" and stated that there was a need for a "well-developed scale to measure online trust that is specifically tailored to the business-to-consumer e-commerce environment" (p. 52). Therefore, we decided to develop a semantic differential that addresses these problems, and which also possesses certain advantages over Likert scales.

1.3. Advantages of semantic differentials

Semantic differentials function by presenting respondents with a set of bipolar items consisting of a pair of antonyms. This provides semantic differentials with specific advantages over the more common Likert-style questionnaires. Respondents to Likert-type scales can only indicate the extent to which they agree or disagree with a specific statement. Hence, a respondent selecting "strongly disagree" does not necessarily imply agreement with the opposite of the item (Chin et al., 2008). Conversely, the format of semantic differentials enables respondents to express their opinion about a concept more fully; that is, ranging from the negative polar to the positive polar. Another advantage is that semantic differentials can reduce the acquiescence bias sometimes provoked by Likert-type scales (Friborg et al., 2006).

The acquiescence bias refers to a category of response biases indicating that respondents have a tendency to agree with all items, or indicating a positive connotation (Friborg et al., 2006). It has also been demonstrated that semantic differentials outperform Likert-based scaling in robustness (Hawkins et al., 1974), reliability (Wirtz and Lee, 2003), and validity (Van Auken and Barry, 1995). Furthermore, semantic differentials function effectively as a short-form scale format, which reduces survey completion time (Chin et al., 2008). Finally, the literature suggests that this format is appropriate when measuring complex and multidimensional constructs (Verhagen et al., 2015).

2. Development and validation strategy

The development and validation followed the framework described by Verhagen et al. (2015). In the first step, relevant literature and existing scales were reviewed to develop a sample of bipolar scales reflecting the underlying concepts of Benevolence, Integrity, and Competence. In the second step, linguistic and psychological bipolarity were established through an extensive review by 18 experts. These scale anchors should function as linguistic and psychological antonyms in relation to the concept being measured. After these two steps, a first study was conducted to reduce the item pool and establish the structural validity (dimensionality) of the scale. We recruited 601 participants in order to conduct an exploratory factor analysis and to investigate correlations of the TrustDiff with related constructs. This served as an initial test of discriminant and convergent validity. A second study, with 312 participants, was conducted to test the measurement model with a confirmatory factor analysis, involving various types of interactive technology. A third study was set up as an experiment with 252 participants, where trust-related elements of a website were actively manipulated to test criterion validity.

3. Item Pool Development and Review

3.1. Item pool

The literature review identified several relevant trust questionnaires and these were used as a basis to develop an initial item. Key adjectives within sentences of existing question-

naires were extracted (Bhattacherjee, 2002; Bart et al., 2005; Cho, 2006; Corbitt et al., 2003; Flavián et al., 2006; Gefen, 2002; Gefen et al., 2003; Hong and Cho, 2011; Jian et al., 2000; Koufaris and Hampton-Sosa, 2004; Lu et al., 2012; McCroskey and Teven, 1999; Pavlou and Gefen, 2004; Rieser and Bernhard, 2016). Forty-three unique adjectives were identified, several of which appeared multiple times in the literature. In a next step, possible antonyms were selected with the help of dictionaries (www.merriam-webster.com, www.thesaurus.com, www.leo.org) and near-duplicates were removed. This process resulted in 28 positive adjectives with up to three different antonyms each.

3.2. Expert review

An item-sort task as well as a test for linguistic and psychological bipolarity were performed by an expert panel ($N = 18$) of trained psychologists and user experienced researchers, using an online survey. These experts assigned each of the 28 adjectives to one of the three dimensions of trust. Adjectives assigned to the correct dimension by fewer than 13 participants were excluded (Howard and Melloy, 2016). For each of the remaining adjectives, the best fitting antonym with the highest agreement was chosen, resulting in an initial item pool of 20 items (refer to Table 1).

4. Study 1

The goal of Study 1 was to reduce the over-representative item pool by employing factor analysis and to test the convergent and discriminant validity of the scale.

4.1. Method

Participants. A total of 714 participants completed the online survey successfully. Responses were excluded from the final data set according to the following criteria: first, if the response time of the participant was under 150 seconds; second, if a repeated response pattern (e.g. crossing only the middle response option for a specific questionnaire) was detected; third, if by the end of the survey the participants themselves started to ignore the data for the final data analysis. After data exclusion, responses from 601 participants (42% women, 58%

Table 1: Items of the trust questionnaire examined in Study 1.

	Item	M	SD	Mdn	S	K
Benevolence						
BEN1	ignoring – caring	4.49	1.241	4	−0.04	−0.51
BEN2	malicious – benevolent	4.49	1.253	4	−0.05	−0.28
BEN3	rude – cordial	5.08	1.158	5	−0.27	−0.24
BEN4	insensitive – sensitive	4.32	1.202	4	−0.03	0.18
BEN5	inconsiderate – empathic	4.52	1.221	4	−0.11	−0.03
Integrity						
INT1	dishonest – honest	4.82	1.356	5	−0.46	−0.13
INT2	insincere – sincere	4.75	1.304	5	−0.45	0.07
INT3	dishonorable – honorable	4.62	1.333	5	−0.22	−0.34
INT4	unbelievable – believable	5.08	1.376	5	−0.71	0.18
INT5	untruthful – truthful	4.93	1.364	5	−0.40	−0.36
INT6	fraudulent – credible	5.06	1.432	5	−0.58	−0.26
Competence						
COM1	clueless – knowledgeable	5.56	1.169	6	−0.91	1.04
COM2	incompetent – competent	5.51	1.211	6	−0.75	0.33
COM3	unskilled – skillful	5.39	1.178	5	−0.59	0.15
COM4	unqualified – proficient	5.39	1.193	6	−0.70	0.45
COM5	incapable – capable	5.55	1.196	6	−0.78	0.58
COM6	uninformed – informed	5.48	1.204	6	−0.65	0.26
COM7	inexperienced – experienced	5.60	1.221	6	−0.89	0.62
COM8	ineffective – effective	5.51	1.244	6	−0.88	0.66
COM9	inept – resourceful	5.43	1.225	6	−0.78	0.53

Note. M = Mean, SD = Standard deviation, Mdn = Median, S = Skew, K = Kurtosis. $N = 601$.

men, Mean age = 38 years, age range: 18 − 84) remained. Recruitment took place on Amazon Mechanical Turk. Participants were reimbursed for their participation with $0.60. Only workers from Amazon Mechanical Turk living in the United States were eligible to participate in the survey.

Procedure and Materials. Participants were asked to perform two tasks on one of two randomly assigned websites. The first group received a link to an online shop (`http://www.crazysales.com.au`), where they were asked to find a product to their liking and to inform themselves about the return policy of the company. The second group was given a link to a website (`http://www.sunshineloans.com.au`) specializing in small loans (refer to section 4.1). While using this website, the participants were asked to inform themselves about loan costs and whether or not security was required when applying for a loan. These two websites were chosen in order to assess trust in a realistic setting. The websites were selected by considering both the website traffic and the website ranking (data from www.alexa.com and www.similarweb.com) in the United States, since the target audience of the survey was inhabitants of the United States and the aim was to select relatively unknown websites to prevent any bias from previous experience. Upon returning to the survey, participants were asked to rate the website according to trust (TrustDiff and a Likert-type Trust scale), usability, and visual aesthetics. Finally, general demographic questions were presented.

4.2. Measures

All items from the questionnaires below were presented in random order within their own subsection of the survey. All measures consisted of seven-point Likert-type scales ranging from 1 (strongly disagree) to 7 (strongly agree), unless otherwise noted.

TrustDiff. The 20 item pairs of the initial version of the TrustDiff were presented as a semantic differential with seven steps between the antonym pairs. Seven steps were chosen because this corresponded to the commonly used seven-point Likert scale and because it has been successfully applied in other semantic differentials (e.g., Hassenzahl et al. (2003)).

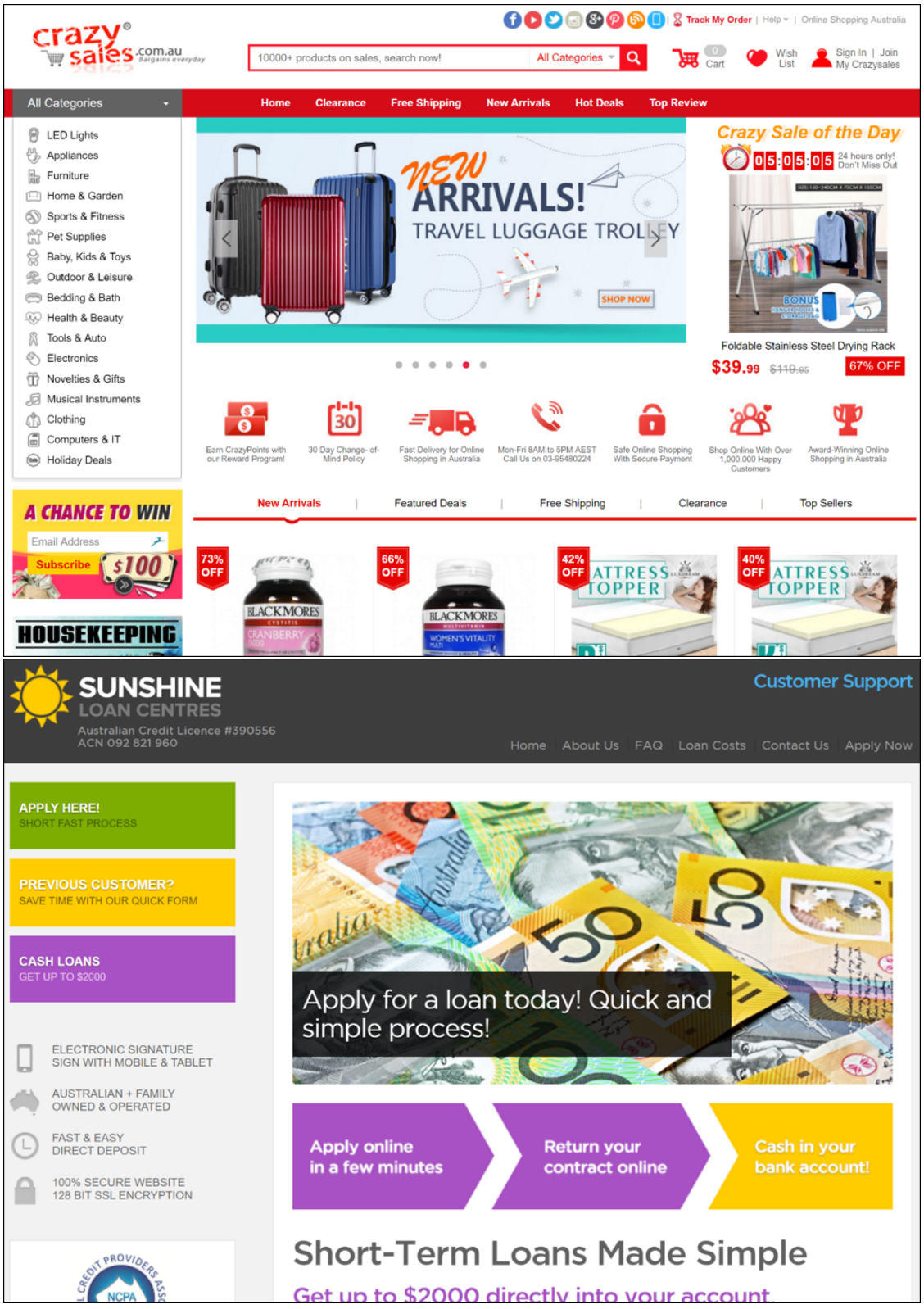


Figure 1: Screenshots from www.crazysales.com.au and www.sunshineloans.com.au

Participants were instructed to rate the website owner ("Please rate the website owner on the following dimensions").

Convergent trust scale. In order to assess the convergent validity, the 15 items of the trust questionnaire developed by Flavián et al. (2006) were included in the survey. Just as the TrustDiff, this Likert-type scale measures trust with three subscales, Benevolence, Integrity and Competence. Slight modifications to the items' declarative statements were made to better fit the measured website. The scale showed excellent internal consistency: benevolence ($\alpha = .90$), integrity ($\alpha = .90$), and competence ($\alpha = .90$).

Visual aesthetics. The discriminant validity of visual aesthetics was assessed using 18 items of the VisAWI (Moshagen and Thielsch, 2010). In an effort to keep the analysis simple, all items were average in an overall aesthetics score. The internal consistency of this scale was excellent (Cronbach's $\alpha = .96$).

Usability Metric for User Experience. As an additional measure of discriminant validity, usability was measured using the four items of the Usability Metric for User Experience (UMUX) (Finstad, 2010). The internal consistency of the scale was good (Cronbach's $\alpha = .87$).

4.3. Results

The full set of $N = 601$ was considered for the item analysis and exploratory factor analysis. A two-samples Kolmogorov-Smirnov test was conducted to make sure that the distributions in the data sets from each website did not differ significantly ($D = 0.090, p = .169$). The item analysis and reduction process followed three steps. First, the distribution statistics for each item were analyzed (see Table 1). Three items (COM1, COM7, COM8) showed a slight negative skew, suggesting a ceiling effect. For this reason and since competence was measured using many items (9), they were excluded from further analysis.

Second, an exploratory factor analysis was conducted on the 17 remaining items with oblique rotation, since factors were expected to be correlated. The Kaiser-Meyer-Olkin measure verified the sampling adequacy of the analysis, KMO = .97 ('marvelous' according

to Hutcheson and Sofroniou (1999)), and all KMO values for individual items were greater than .95, which was well above the acceptable limit of .5 (Field, 2013). The Bartlett's Test of sphericity, which tests the overall significance of all correlations within the correlation matrix, was significant ($\chi^2(136) = 9533.923, p < .001$), suggesting that using an exploratory factor analysis was appropriate. In an initial analysis of the eigenvalues, only two factors had eigenvalues over Kaiser's criterion of 1. However, the parallel analysis and the screen plot suggested three factors that in combination explained 61% of the variance. The exploratory factor analysis was performed using three factors, as this solution was in line with the theoretical model of three subcomponents of trust. After the first exploratory analysis, a total of three items (BEN3, INT2, and INT3) was eliminated because they did not contribute to the factor structure and failed to meet the minimum criteria (Howard, 2016) of having a primary factor loading of .4 or above, and no cross-loading of .3 or above (see Table 2).

A second exploratory factor analysis of the remaining 14 items, again with minres and oblimin rotation, was conducted. The three factors explained 74% of the variance. All items had primary loadings above .5 and load with their corresponding factor. The factor loadings are presented in Table 3 and the correlations between the factors are presented in Table 4. Finally, the reliability of each subscale was analyzed. Benevolence ($\alpha = .89$), integrity ($\alpha = .95$) and competence ($\alpha = .93$) showed high internal consistency. No substantial increase in Cronbach's alpha for any of the scales could have been achieved by eliminating more items.

4.4. Convergent and discriminant validity

In order to assess convergent and discriminant validity, the correlation between the TrustDiff and related measures was explored. Table 5 shows that the TrustDiff correlated strongly ($r = .68$) with the trust questionnaire adapted from Flavián et al. (2006), indicating convergent validity. The TrustDiff scale was found to correlate with visual aesthetics as well as usability ($r = .46$ and $r = .50$ respectively). Interestingly, the subscale Benevolence was less strongly related to visual aesthetics and usability than the other subscales ($r = .33$ and $r = .34$ compared to correlations in the range of $.41 - .57$).

Table 2: Rotated pattern matrix of the exploratory factor analysis in Study 1.

	Factor loadings			
Item	Benevolence	Integrity	Competence	h2
BEN1: ignoring − caring	**.774**	.079	.071	.767
BEN2: malicious − benevolent	**.616**	.179	.054	.629
BEN3: rude − cordial	**.446**	.070	**.319**	.530
BEN4: insensitive − sensitive	**.848**	−.018	−.005	.691
BEN5: inconsiderate − empathic	**.860**	.000	.016	.753
INT1: dishonest − honest	.143	**.849**	−.076	.830
INT2: insincere − sincere	**.401**	**.508**	.012	.741
INT3: dishonorable − honorable	**.472**	**.430**	.042	.764
INT4: unbelievable − believable	.086	**.693**	.082	.671
INT5: untruthful − truthful	−.035	**.732**	.160	.701
INT6: fraudulent − credible	−.035	**.747**	.205	.768
COM2: incompetent − competent	−.047	.126	**.823**	.791
COM3: unskilled − skillful	.099	−.084	**.846**	.707
COM4: unqualified − proficient	−.004	.067	**.828**	.763
COM5: incapable − capable	−.062	.114	**.841**	.793
COM6: uninformed − informed	−.027	.076	**.832**	.760
COM9: inept − resourceful	.125	−.145	**.868**	.699
Eigenvalues	1.98	0.73	10.46	
% of variance	18	17	26	

Note. Exploratory factor analysis with minres and oblimin.

Factor loadings above .3 are marked in bold. N = 601.

Three factors explain 61% of the total variance. h2 = Communality, $N = 601$.

Table 3: Results of the second exploratory factor analysis in Study 1.

	Factor loadings			
Item	Benevolence	Integrity	Competence	h2
BEN1: ignoring − caring	**.785**	.081	.059	.779
BEN2: malicious − benevolent	**.605**	.174	.058	.611
BEN4: insensitive − sensitive	**.825**	.005	−.014	.675
BEN5: inconsiderate − empathic	**.903**	−.025	.009	.790
INT1: dishonest − honest	.143	**.877**	−.113	.834
INT4: unbelievable − believable	.074	**.709**	.060	.657
INT5: untruthful − truthful	−.011	**.762**	.121	.714
INT6: fraudulent − credible	−.030	**.770**	.173	.774
COM2: incompetent − competent	−.040	.121	**.822**	.793
COM3: unskilled − skillful	.087	−.065	**.836**	.700
COM4: unqualified − proficient	−.002	.065	**.827**	.762
COM5: incapable − capable	−.051	.097	**.847**	.795
COM6: uninformed − informed	−.014	.055	**.841**	.764
COM9: inept − resourceful	.114	−.142	**.871**	.693
Eigenvalues	0.70	1.80	8.62	
% of variance	18	18	38	
α	.90	.92	.95	

Note. Three factors explain 74% of the total variance.

Factor loadings above .3 are marked in bold. N = 601.

Table 4: Correlations between the factors extracted in Study 1.

Factor	Benevolence	Integrity	Competence
Benevolence	–		
Integrity	.76	–	
Competence	.52	.72	–

Note. $N = 601$.

Table 5: Descriptive statistics and Pearson correlations of measures in Study 1.

	M	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
TrustDiff												
1. Benevolence	4.45	1.08	–									
2. Integrity	4.97	1.24	.74	–								
3. Competence	5.46	1.07	.55	.72	–							
4. Total	4.96	1.00	.86	.94	.85	–						
Flavián et al. (2006)												
5. Benevolence	4.79	1.21	.60	.66	.54	.68	–					
6. Integrity	4.83	1.18	.54	.69	.54	.67	.86	–				
7. Competence	5.24	1.18	.36	.47	.57	.53	.75	.78	–			
8. Total	4.95	1.11	.54	.65	.59	.68	.93	.95	.91	–		
9. VisAWI	4.74	1.22	.33	.41	.47	.46	.46	.48	.49	.51	–	
10. UMUX	5.42	1.21	.34	.47	.53	.50	.48	.51	.55	.55	.75	–

Note. $N = 601$. All correlations are significant $p < .001$

4.5. Discussion

In Study 1, 14 items measuring three related subcomponents of trust were identified. Analysis of correlations with related measures, such as an existing rating scale, offered a first test of convergent validity. Comparatively low correlations of the TrustDiff with visual aesthetics and usability indicated discriminant validity. The results of the second exploratory factor analysis supported a three dimensional measure with high reliability and good psychometric properties. The measurement model of the TrustDiff was tested and refined in Study 2. The ability of the final TrustDiff to differentiate between two manipulated websites was investigated in Study 3.

5. Study 2

5.1. Method

Procedure and Measures. As part of a larger study, participants were asked to name a single interactive technology they used frequently. Participants indicated how often they had used this particular technology over the last 14 days. The rest of the online survey focused on this particular technology, and the 14 items of the TrustDiff in Study 1 were included. As in Study 1, the word pairs were presented in random order.

Participants. A total of 315 participants from the United States completed the relevant part of the survey on Mechanical Turk. Three participants had to be excluded because they indicated that we should not use their data, resulting in a final sample of $N = 312$ (55% women, 44% men, 1% other or not disclosed; mean age = 37.6 years, age range: $18 - 76$).

Type of technology and frequency of use. The most frequently mentioned technology was Facebook (42.7%), followed by other social media (Twitter 7.4%, Instagram 7.1%, YouTube 3.5%), Fitbit (3.2%), Microsoft Word or Excel (2.6%, 1.9%) and various other technologies, among them Mechanical Turk, web browser, Amazon Alexa, digital games, and mobile apps. The vast majority of participants indicated that they used the technology multiple times a day (84.6%). Almost 44% indicated that they used the technology six or more times a day.

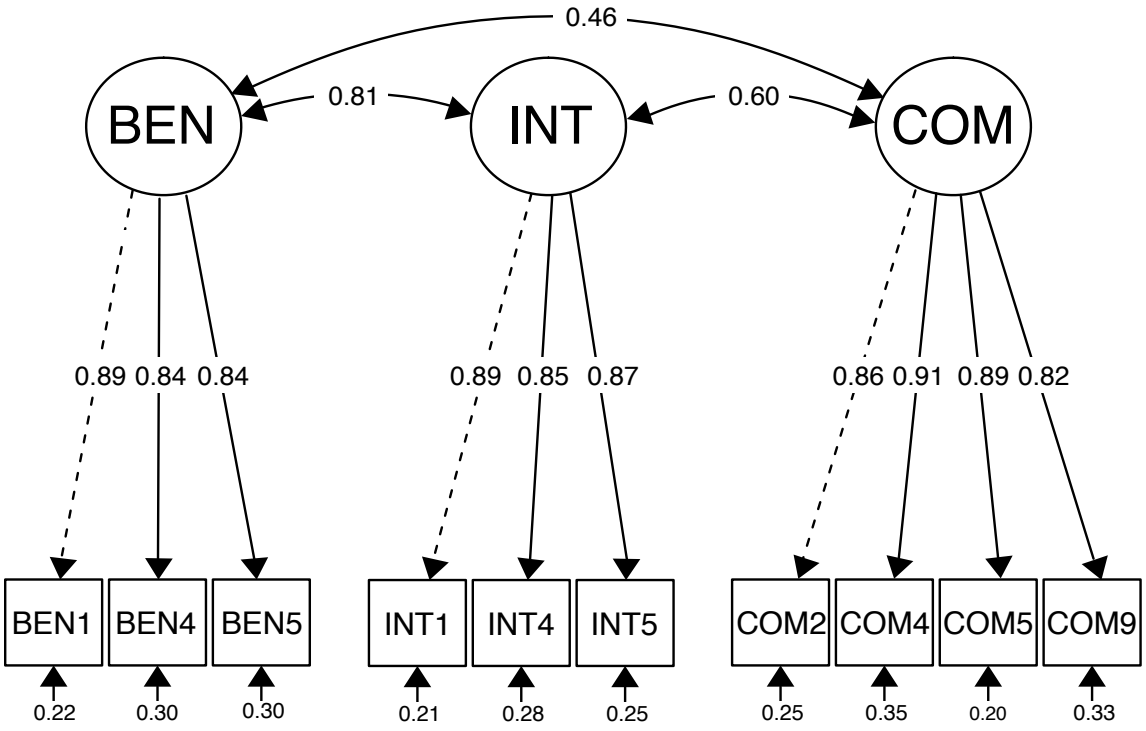


Figure 2: Measurment model of the TrustDiff in Study 2 with standardized loadings. Dotted lines indicate loadings that were constrained to one [$\chi^2(32) = 32.500$, $p = .442$, $\chi^2/df = 1.02$, $CFI = 1.000$, $SRMR = .027$, $RMSEA = .007$, $PCLOSE = .996$]

5.2. Results and Discussion

As a test of the three-dimensional factor structure, a confirmatory factor analysis was conducted using the lavaan package (0.5-23.1097) for R. All items were specified to load on their designated factor, and the loading of the first item was constrained to one. Multivariate normality was not given (Mardia tests: $\chi^2_s = 2474.4, p < .001$; $Z_k = 50.6, p < .001$), therefore we used a robust maximum likelihood estimation method with Huber-White standard errors and a Yuan-Bentler based scaled test statistic. Results of the CFA including all 14 items suggested that the proposed model adequately but not perfectly fitted the data [$\chi^2(74) = 140.530$, $p < .001$, $\chi^2/df = 1.89$, $CFI = .971$, $SRMR = .047$, $RMSEA = .054$, $PCLOSE = .279$]. All loadings of the latent factors on their designated items exceeded .80 except for item BEN2 (.67). Investigation of modification indices suggested covariance

between items COM3 and COM4 as well as between COM3 and COM5, and a loading of Competence on IMT4 to improve model fit. However, since the goal was to create an economic scale for user trust with three subscales, certain items were removed instead of allowing cross-loadings for a better model fit. Thus, item BEN2 (malicious benevolent) was removed because of low loadings of the Benevolence factor, INT4 (unbelievable believable) was removed to reduce a possible influence by Competence on Integrity, and Item COM5 (uninformed informed) was removed, primarily on theoretical grounds. The aspect of how informed a vendor of a product was seemed to be less related to other aspects of competence such as capability, qualifications and resources. The item COM3 (unskilled skillful) was removed because it had too much statistical and theoretical overlap with item COM4 (unqualified proficient).

The final scale was reduced to 10 items, measuring the three related but distinct dimensions, and showed excellent psychometric properties [$\chi^2(32) = 32.500$, $p = .442$, $\chi^2/df = 1.02$, $CFI = 1.000$, $SRMR = .027$, $RMSEA = .007$, $PCLOSE = .996$]. Descriptive statistics of the final 10-item TrustDiff are provided in Table 6 and the measurement model is shown in Figure 2. Internal consistency of the three subscales was high ($\alpha_{Ben} = .85$, $\alpha_{Int} = .90$, $\alpha_{Com} = .91$).

The results of these two confirmatory factor analyses showed that the questionnaire could be improved and shortened without losing reliability. The final model for the 10-item TrustDiff presented an excellent fit with high internal consistency. In the next step, an experiment was conducted to investigate the criterion validity of the scale.

6. Study 3

The goal of Study 3 was to test whether the TrustDiff was able to differentiate between two websites when their trust-related features were manipulated.

6.1. Method

Procedure and Materials. As part of a larger research project, but unrelated to Study 1 or Study 2, participants were asked to rate a mock online shop based on a provided screenshot.

Table 6: Descriptive statistics and Pearson correlations of all items for the final TrustDiff in Study 2.

	M	SD	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. BEN1: ignoring − caring	4.43	1.39	–									
2. BEN4: insensitive − sensitive	4.27	1.38	0.74	–								
3. BEN5: inconsiderate − empathic	4.46	1.37	0.74	0.71	–							
4. INT1: dishonest − honest	4.78	1.47	0.64	0.59	0.61	–						
5. INT4: unbelievable − believable	5.06	1.50	0.59	0.56	0.56	0.76	–					
6. INT5: untruthful − truthful	4.68	1.45	0.65	0.59	0.59	0.78	0.72	–				
7. COM2: incompetent − competent	5.79	1.24	0.34	0.29	0.30	0.40	0.50	0.46	–			
8. COM4: unqualified − proficient	5.64	1.31	0.34	0.29	0.32	0.37	0.48	0.39	0.69	–		
9. COM5: incapable − capable	5.81	1.29	0.40	0.34	0.41	0.47	0.52	0.49	0.77	0.73	–	
10.COM9: inept − resourceful	5.79	1.34	0.36	0.27	0.30	0.36	0.48	0.42	0.72	0.68	0.72	–

Note. $N = 312$. All correlations are significant $p < .001$

The participants were randomly assigned to one of two groups. The first group was presented with a screenshot of an online shop that included several trust-supporting elements (high trust), while the second group received a screenshot of an online shop that was lacking any trust-supporting elements (neutral) (see Figure 3). Graphic design, structure design, content design, and social-cue design elements were manipulated (see Table 7) according to the elements identified by Seckler et al. (2015) and Wang and Emurian (2005).

After examining the website screenshot for at least four seconds, participants were asked to fill in the TrustDiff, the Likert-type scale for trust by Flavián et al. (2006), and to rate the visual appeal and perceived usability of the website. All measures were presented as in Study 1. Data collected for this part were used to assess whether the TrustDiff could differentiate between high and neutral trustworthiness.

Participants. A total of 394 participants from the United States completed the relevant part of the survey on CrowdFlower. Data were cleaned with two attention check items, which reduced the sample size to 258. Six additional participants were excluded because they indicated not to use their data, resulting in a final sample of $N = 252$ (71% women, 28% men, 1% other or not disclosed; mean age = 39 years, age range: $18 - 78$).

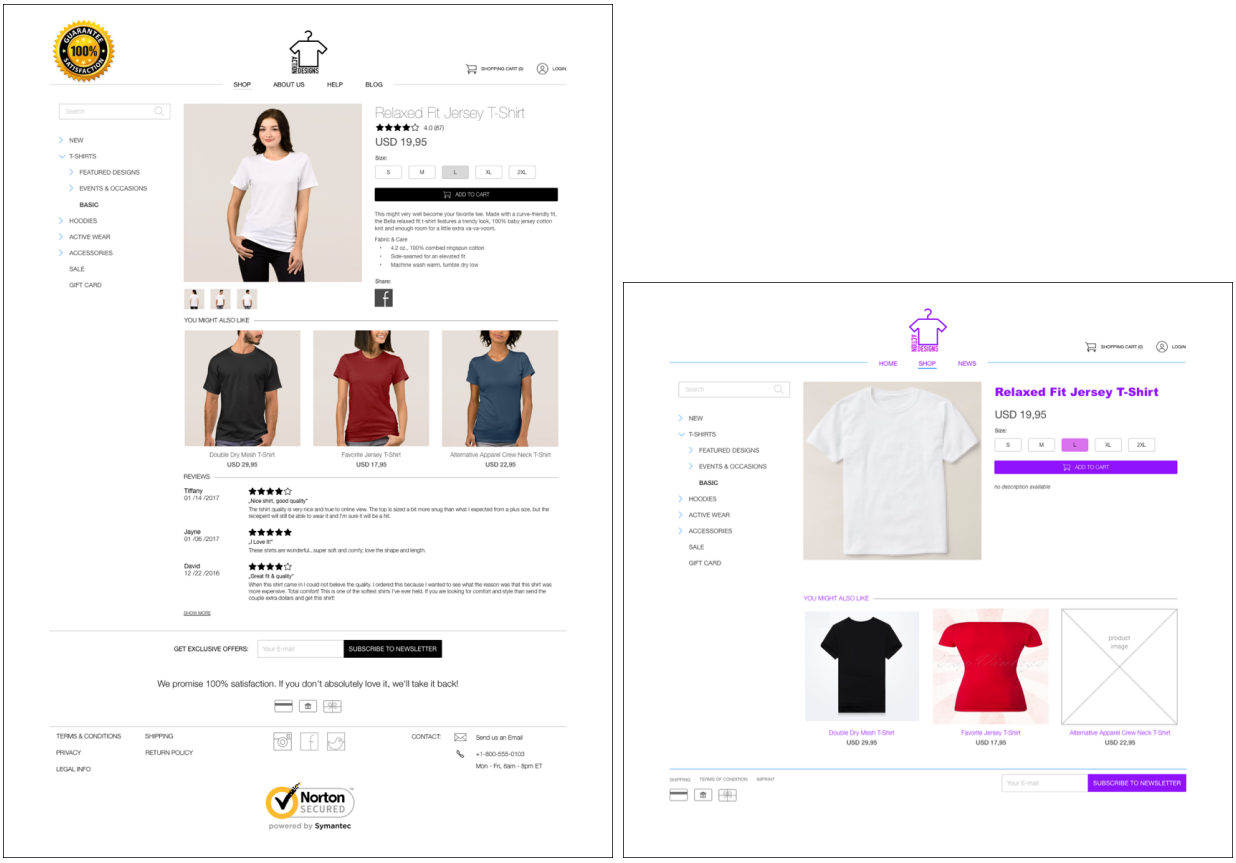


Figure 3: Mock online shop with trust-related features (left) and without (right) used in the experiment in Study 2. *Note to reviewers: a high-resolution image is included as a separate file at the end of this manuscript*

Table 7: Overview of the dimensions and respective features manipulated in the mock online shop.

Dimension	High-Trust	Neutral
Graphic design	Muted colours, high contrast	Bright colours, lower contrast
	Well-chosen and well-shot photographs	Inconsistent and missing photographs
Content design	Satisfaction guarantee	No satisfaction guarantee
	Links to more information in the footer, clearly readable	Hard to read or lacking information
	Link to the privacy policy	No link visible
	Seals of approval or third-party certificate	No seals of approval or third-party certificates
	Comprehensive, correct, and current product information	No product information
Social-cue design	Contact information for customer service in the footer	No contact information
	Users' reviews visible	Lack of users' reviews

Measures. The 10 word-pairs of the final TrustDiff were included together with 15 items from the Trust scale developed by Flavián et al. (2006) and 18 items from the VisAWI measure for visual aesthetics Moshagen and Thielsch (2010)(Cronbach's $\alpha = .95$). Unlike Study 1, only the overall score on the Trust scale by Flavián et al. (2006) was included in the analysis (Cronbach's $\alpha = .96$). All three subscales of the TrustDiff showed excellent internal consistency ($\alpha_{Ben} = .86$, $\alpha_{Int} = .90$, $\alpha_{Com} = .94$). As in Study 1 and Study 2, seven-point scales were employed and the items were presented randomly.

6.2. Results

On average, participants viewed the websites for 1.47 minutes ($SD = 1.4$, min = 13.8 seconds, max = 14.08 minutes). No significant differences in viewing time (log-transformed) were observed between the conditions, $t(246.88) = 0.073065$, $p = 0.9418$. All measures deviated significantly from normal distribution, therefore Welch's two samples t-test and robust Wilcoxon rank-sum tests were conducted. Both tests led to the same conclusions for all measures, with the result that we decided to list only the results of the Welch's t-test. Criterion validity was investigated by comparing the high trust condition with the

Table 8: Descriptive statistics and results of Welch's two samples t-test as an assessment of criterion validity of the TrustDiff.

		High trust ($n = 128$)		Neutral ($n = 124$)					
		M	SD	M	SD	t	df	p	d
TrustDiff	Benevolence	5.01	0.962	4.28	1.075	5.681	245.0	$< .001$	0.72
	Integrity	5.48	0.993	4.75	1.199	5.210	238.7	$< .001$	0.66
	Competence	5.58	1.011	4.47	1.455	6.989	218.6	$< .001$	0.89
	Total	5.37	0.899	4.50	1.176	6.577	230.1	$< .001$	0.84
Trust		5.11	0.947	4.20	1.254	6.470	228.8	$< .001$	0.82
VisAWI		4.91	1.094	3.91	1.167	7.037	247.7	$< .001$	0.89

Note. Total $N = 252$.

neutral condition. As presented in Table 8, Welch's two samples t-tests yielded significant differences between the conditions for all subscales of the TrustDiff and the total score ($t(230.1) = 6.577$, $p < .001$, $d = 0.84$). The Likert-type scale for trust Flavián et al. (2006) also showed a significant difference between the two conditions ($t(228.8) = 6.470$, $p < .001$, $d = 0.82$). The difference between the two websites was even more pronounced for aesthetics, which was generally rated lower by the participants ($t(247.7) = 7.037$, $p < .001$, $d = 0.89$).

7. General Discussion

The aim of this project was to develop and validate a scale measuring trust in online contexts, using a semantic differential. Scale construction is an important step in confirmatory research because the quality of a measurement scale determines the extent to which empirical results are meaningful and accurate (Bhattacherjee, 2002).

The main contribution of the TrustDiff is two-fold: first, the semantic differential ensures a broad applicability for measuring user trust on the web. As discussed earlier, the majority of existing trust questionnaires make use of Likert-type items, which are mostly tailored to the specific website measured. This makes it difficult to use these questionnaires in other

research contexts (e.g., Lu et al., 2012; McKnight et al., 2002). The pairs of antonyms used in the TrustDiff, however, comprise adjectives which generally fit any context related to user trust on the web. Second, each item of the TrustDiff contains two words only (one item-pair), namely two contrary adjectives, which are easier to translate into other languages than full sentences. The declarative statements used in Likert-scale items from other trust scales (e.g., Bhattacherjee, 2002; Cho, 2006; Flavián et al., 2006; Gefen, 2002; McKnight et al., 2002) on the other hand are often complex and time-consuming to translate. Taken together, the advantages of the TrustDiff over other trust scales is its broader and easier applicability in different contexts and languages, keeping the possible loss of reliability and validity to a minimum level, and its ability to measure different manifestations of trust from a negative to a positive pole in one scale. International firms whose online-services are available across several countries and different languages might profit from a universally applicable trust scale. A company may lose a great deal if it fails to assess consumer trust in its services, especially when revenue structure depends on frequent and continuous user transactions. Early identification of users with low trust levels may help to ensure their retention by targeting them specifically with specialized interventions.

Based on the existing literature, 28 positive adjectives, with up to three antonyms each were generated for the three dimensions of trust (Benevolence, Integrity, and Competence). These items were tested for appropriate linguistic and psychological bipolarity by an expert panel and reduced to 20 item pairs. Results from factor analysis in Study 1 (N = 601) suggested a 14-item scale measuring three distinct but related dimensions of trust. The trust dimensions of the 14-item TrustDiff were relatively highly correlated with a Likert-type trust scale and substantially correlated, although this was less pronounced, with perceived usability and aesthetics. In Study 2, the 14-item questionnaire measurement model was tested with 312 participants who rated various frequently used technologies. Results of a confirmatory factor analysis suggested several avenues for improvement, which resulted in a 10-item scale for trust with good psychometric properties. The results of Study 3 showed that the TrustDiff was sensitive to websites with differences in trust-related features. The rating differences between the two websites were between d = 0.66 and 0.89, commonly interpreted

as between moderate to large (Cohen, 1977). Compared to existing questionnaires that are content specific (e.g., McKnight et al., 2002) or developed in other languages (e.g., Flavián et al., 2006), the TrustDiff can be applied in various contexts and has been tested with English-speaking participants. From a practitioners standpoint, the 10-item TrustDiff can be applied without modifications to assess customers' levels of trust in an enterprise or service, and may be translated easily into other languages.

The three studies presented here entail an initially thorough validation of the TrustDiff. Although the scale offers promising psychometric properties, the TrustDiff requires further testing with various products and services in different contexts. However, the 10-item scale showed very good psychometric properties with a wide variety of technologies in Study 2. The structure of the TrustDiff found in Study 2 needs to be replicated in different cultural contexts and with languages other than English. For this task, a semantic differential is ideal, as less translation effort is needed than with traditional Likert-type scales. However, it is still essential to establish psychometric bipolarity and structural validity in other languages.

The TrustDiff could be used to investigate how different web design elements relate to the different dimensions of trust or distrust, since the present questionnaire represents the construct trust from a negative to a positive pole on three subscales. Furthermore, in order to build a comprehensive picture of users trust and trust-related behaviors, the TrustDiff could be combined with measures of trust in a technology. Trust in a technology has been found to be related to the intention to explore and use more features of that particular technology (McKnight et al., 2011). This vendor-technology trust distinction could be particularly helpful in better understanding their relative influence on the adoption of a technology, post-adoption use and the abandonment of a technology. Ultimately, researchers could investigate the predictive power of the TrustDiff with regard to the trust-related behavior of users and how this may relate to antecedents of trust. For instance, interface language quality, which is a major issue in multilingual software projects (e.g., Bargas-Avila and Brühlmann, 2016) could influence users trust in vendors. In addition, the wording of the TrustDiff is not exclusive to the web context since many of the items might have face-validity in other settings. For instance, the validity of the TrustDiff could be investigated

in areas of interpersonal trust or off-line buyer-seller relationships. No less promising would be an attempt to discover profiles based on users' responses on the scale. This may allow researchers and practitioners to design and evaluate trust-related interventions targeted at specific subgroups.

8. Conclusion

We present the development and validation of a semantic differential that helps to evaluate users' trust, and which has the potential to serve as a tool to investigate how user trust emerges. The development and validation followed best practices and the scale is readily applicable to a variety of research questions. The TrustDiff was tested with over 1000 participants and showed good psychometric properties and high reliability. The semantic differential is easy to use and easy to translate and thus a viable alternative to existing Likert scale format questionnaires on user trust.

9. Acknowledgements

Special thanks to Elisa Mekler. This work has been approved by the Institutional Review Board of the Faculty of Psychology, University of Basel under the numbers D-003-17 and M-003-17.

10. References

L. V. Casaló, C. Flavián, M. Guinalíu, The role of security, privacy, usability and reputation in the development of online banking, Online Information Review 31 (2007) 583–603.

J. W. Driscoll, Trust and participation in organizational decision making as predictors of satisfaction, Academy of Management Journal 21 (1978) 44–56.

C. Moorman, R. Deshpande, G. Zaltman, Factors affecting trust in market research relationships, Journal of Marketing (1993) 81–101.

J. B. Rotter, A new scale for the measurement of interpersonal trust, Journal of Personality 35 (1967) 651–665.

R. J. Lewicki, C. Brinseld, Measuring trust beliefs and behaviours, in: F. Lyon, G. Mollering, M. N. K. Saunders (Eds.), Handbook of research methods on trust, Edward Elgar Publishing, Cheltenham, UK, 2012, pp. 29–39.

L. van der Werff, C. Real, T. Lynn, Individual trust and the internet, in: R. H. Searle, A.-M. I. Nienaber, S. B. Sitkin (Eds.), The Routledge Companion to Trust, Routledge, Oxford, UK, 2018.

Y. D. Wang, H. H. Emurian, An overview of online trust: Concepts, elements, and implications, Computers in Human Behavior 21 (2005) 105–125.

A. Bhattacherjee, Individual trust in online firms: Scale development and initial test, Journal of Management Information Systems 19 (2002) 211–241.

J. Cho, The mechanism of trust and distrust formation and their relational outcomes, Journal of Retailing 82 (2006) 25–35.

C. Flavián, M. Guinalíu, R. Gurrea, The role played by perceived usability, satisfaction and consumer trust on website loyalty, Information & Management 43 (2006) 1–14.

D. Gefen, Reflections on the dimensions of trust and trustworthiness among online consumers, ACM Sigmis Database 33 (2002) 38–53.

D. H. McKnight, V. Choudhury, C. Kacmar, The impact of initial consumer trust on intentions to transact with a web site: A trust building model, The Journal of Strategic Information Systems 11 (2002) 297–323.

Y. Kim, R. A. Peterson, A meta-analysis of online trust relationships in e-commerce, Journal of Interactive Marketing 38 (2017) 44–54.

T. Verhagen, B. van Den Hooff, S. Meents, Toward a better use of the semantic differential in is research: An integrative framework of suggested action., Journal of the Association for Information Systems 16 (2015) 108–143.

S. C. Chen, G. S. Dhillon, Interpreting dimensions of consumer trust in e-commerce, Information Technology and Management 4 (2003) 303–318.

R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of Management Review 20 (1995) 709–734.

Y. Bart, V. Shankar, F. Sultan, G. L. Urban, Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study, Journal of Marketing 69 (2005) 133–152.

B. J. Corbitt, T. Thanasankit, H. Yi, Trust and e-commerce: A study of consumer perceptions, Electronic Commerce Research and Applications 2 (2003) 203–215.

M. K. Lee, E. Turban, A trust model for consumer internet shopping, International Journal of Electronic Commerce 6 (2001) 75–91.

S. L. Jarvenpaa, N. Tractinsky, L. Saarinen, Consumer trust in an internet store: A cross-cultural validation, Journal of Computer-Mediated Communication 5 (1999) 0–0.

D. H. McKnight, V. Choudhury, C. Kacmar, Developing and validating trust measures for e-commerce: An integrative typology, Information Systems Research 13 (2002) 334–359.

P. A. Pavlou, D. Gefen, Building effective online marketplaces with institution-based trust, Information Systems Research 15 (2004) 37–59.

J. Lu, L. Wang, L. A. Hayes, How do technology readiness, platform functionality and trust influence C2C user satisfaction?, Journal of Electronic Commerce Research 13 (2012) 50–69.

W. W. Chin, N. Johnson, A. Schwarz, A fast form approach to measuring technology acceptance and other constructs, MIS Quarterly (2008) 687–703.

O. Friborg, M. Martinussen, J. H. Rosenvinge, Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience, Personality and Individual Differences 40 (2006) 873–884.

D. I. Hawkins, G. Albaum, R. Best, Stapel scale or semantic differential in marketing research?, Journal of Marketing Research 11 (1974) 318–322.

J. Wirtz, M. C. Lee, An examination of the quality and context-specific applicability of commonly used customer satisfaction measures, Journal of Service Research 5 (2003) 345–355.

S. Van Auken, T. E. Barry, An assessment of the trait validity of cognitive age measures, Journal of Consumer Psychology 4 (1995) 107–132.

D. Gefen, E. Karahanna, D. W. Straub, Trust and TAM in online shopping: An integrated model, MIS Quarterly 27 (2003) 51–90.

I. B. Hong, H. Cho, The impact of consumer trust on attitudinal loyalty and purchase intentions in b2c e-marketplaces: Intermediary trust vs. seller trust, International Journal of Information Management 31 (2011) 469–479.

J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated systems, International Journal of Cognitive Ergonomics 4 (2000) 53–71.

M. Koufaris, W. Hampton-Sosa, The development of initial trust in an online company by new customers, Information & Management 41 (2004) 377–397.

J. C. McCroskey, J. J. Teven, Goodwill: A reexamination of the construct and its measurement, Communications Monographs 66 (1999) 90–103.

D. C. Rieser, O. Bernhard, Measuring trust: The simpler the better?, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, New York, NY, USA, 2016, pp. 2940–2946.

M. C. Howard, R. C. Melloy, Evaluating item-sort task methods: The presentation of a new statistical significance formula and methodological best practices, Journal of Business and Psychology 31 (2016) 173–186.

M. Hassenzahl, M. Burmester, F. Koller, Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität, in: Mensch & Computer 2003, Springer, 2003, pp. 187–196.

M. Moshagen, M. T. Thielsch, Facets of visual aesthetics, International Journal of Human-Computer Studies 68 (2010) 689–709.

K. Finstad, The usability metric for user experience, Interacting with Computers 22 (2010) 323 – 327.

G. D. Hutcheson, N. Sofroniou, The multivariate social scientist: Introductory statistics using generalized linear models, Sage, 1999.

A. Field, Discovering statistics using IBM SPSS statistics, Sage, 2013.

M. C. Howard, A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?, International Journal of Human-Computer Interaction 32 (2016) 51–62.

M. Seckler, S. Heinz, S. Forde, A. N. Tuch, K. Opwis, Trust and distrust on the web: User experiences and website characteristics, Computers in Human Behavior 45 (2015) 39—50.

J. Cohen, Statistical Power Analysis for the Behavioral Sciences (Revised Edition), Academic Press, 1977.

D. H. McKnight, M. Carter, J. B. Thatcher, P. F. Clay, Trust in a specific technology: An investigation of its components and measures, ACM Trans. Manage. Inf. Syst. 2 (2011) 12:1–12:25.

J. A. Bargas-Avila, F. Brühlmann, Measuring user rated language quality: Development and validation of the user interface language quality survey (LQS), International Journal of Human-Computer Studies 86 (2016) 1–10.